Towards an Interactive Human-Robot Relationship: Developing a Customized Robot Behavior to Human Profile.

Amir Aly

Towards an Interactive Human-Robot Relationship: Developing a Customized Robot's Behavior to Human's Profile

Thesis Research

Submitted to the Cognitive Robotics and Vision Laboratory
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

ENSTA ParisTech

École Nationale Supérieure de Techniques Avancées

By

Amir ALY

France, December 16, 2014

Thesis Examination Committee

Reviewer:	Prof. Jean-Claude Martin	University of Paris 11 - France
Reviewer:	Prof. Rachid Alami	LAAS CNRS - France
Examiner:	Prof. Angelo Cangelosi	University of Plymouth - England
Examiner:	Prof. Peter-Ford Dominey	INSERM CNRS - France
Examiner:	Prof. Ginevra Castellano	Uppsala University - Sweden
		& University of Birmingham - England
Examiner:	Prof. Nicola Bellotto	University of Lincoln - England
Supervisor:	Prof. Adriana Tapus	ENSTA ParisTech - France

Abstract

Robots become more and more omnipresent in our life and society, and many challenges arise when we try to use them in a social context. This thesis focuses on how to generate an adapted robot's behavior to human's profile so as to enhance the human-robot relationship. This research addresses a wide range of complex problems varying from analyzing and understanding human's emotion and personality to synthesizing a complete synchronized multimodal behavior that combines gestures, speech, and facial expressions. Our methodologies have been examined experimentally with NAO robot from Aldebaran Robotics and ALICE robot from Hanson Robotics.

The first part of this thesis focuses on emotion analysis and discusses its evolutionary nature. The fuzzy nature of emotions imposes a big obstacle in front of defining precise membership criteria for each emotion class. Therefore, fuzzy logic looks appropriate for modeling these complex data sets, as it imitates human logic by using a descriptive and imprecise language in order to cope with fuzzy data. The variation of emotion expressivity through cultures and the difficulty of including many emotion categories inside one database, makes the need for an online recognition system of emotion as a critical issue. A new online fuzzy-based emotion recognition system through prosodic cues was developed in order to detect whether the expressed emotion confirms one of the previously learned emotion clusters, or it constitutes a new cluster (not learned before) that requires a new verbal and/or nonverbal action to be synthesized.

On the other hand, the second part of this thesis focuses on personality traits, which play a major role in human social interaction. Different researches studied the long term effect of the extraversion-introversion personality trait on human's generated multimodal behavior. This trait can, therefore, be used to characterize the combined verbal and nonverbal behavior of a human interacting with a robot so as to allow the robot to adapt its generated multimodal behavior to the interacting human's personality. This behavior adaptation could follow either the similarity attraction principle (i.e., individuals are more attracted by others who have similar personality traits) or the complementarity attraction principle (i.e., individuals are more attracted by others whose personalities are complementary to their own personalities) according to the context of interaction. In this thesis, we examine the effects of the multimodality and unimodality of the generated behavior on interaction, in addition to the similarity attraction principle as it considers the effect of the initial interaction between human and robot on the developing relationship (e.g., friendship), which makes it more appropriate for our interaction context. The detection of human's personality trait as being introverted or extraverted is based on a psycholinguistic analysis of human's speech, upon which the characteristics of the generated robot's speech and gestures are defined.

Last but not least, the third part of this thesis focuses on gesture synthesis. The generation of appropriate head-arm metaphoric gestures does not follow a specific linguistic analysis. It is mainly based on the prosodic cues of human's speech, which correlate firmly with emotion and the dynamic characteristics of metaphoric gestures. The proposed system uses the Coupled Hidden Markov Models (CHMM) that contain two chains for modeling the characteristic curves of the segmented speech and gestures. When a speech-test signal

is present to the trained CHMM, a corresponding set of adapted metaphoric gestures will be synthesized. An experimental study (in which the robot adapts the emotional content of its generated multimodal behavior to the context of interaction) is set for examining the emotional content of the generated robot's metaphoric gestures by human's feedback directly. Besides, we examine the effects of both the generated facial expressions using the expressive face of ALICE robot, and the synthesized emotional speech using the text to speech toolkit (Mary-TTS) on enhancing the expressivity of the robot, in addition to comparing between the effects of the multimodal interaction and the interaction that employs less affective cues on human.

Generally, the research on understanding human's profile and generating an adapted robot's behavior opens the door to other topics that need to be addressed in an elaborate way. These topics include, but not limited to: developing a computational cognitive architecture that can simulate the functionalities of the human brain areas that allow understanding and generating speech and physical actions appropriately to the context of interaction, which constitutes a future research scope for this thesis.

Résumé

Les robots deviennent de plus en plus omniprésents dans la vie et la société, et de nombreux défis surviennent lorsque nous essayons de les utiliser dans un contexte social. Cette thèse porte sur la façon de générer un comportement adapté du robot au profil de l'homme afin d'améliorer la relation homme-robot. Cette recherche aborde un large éventail de problèmes complexes allant de l'analyse et la compréhension de l'émotion et de la personnalité de l'homme à la synthèse d'un comportement multimodal synchronisé qui combine les gestes, la parole, et les expressions faciales. Nos méthodes ont été examinées expérimentalement avec le robot NAO d'Aldebaran Robotics et le robot ALICE de Hanson Robotics.

La première partie de cette thèse porte sur l'analyse de l'émotion. La nature floue des émotions impose un grand obstacle devant la définition des critères d'adhésion précis pour chaque classe d'émotion. Par conséquent, la logique floue semble appropriée pour la modélisation de ces ensembles de données complexes, car elle imite la logique humaine en utilisant un langage descriptif et imprécis pour faire face aux données floues. La variation de l'expressivité de l'émotion à travers les cultures et la difficulté d'inclure de nombreuses catégories d'émotion à l'intérieur d'une base de données, rend le besoin d'un système de reconnaissance en ligne de l'émotion comme un problème critique. Un nouveau système flou à base de reconnaissance des émotions en ligne par des indices prosodiques a été développé afin de détecter si l'émotion exprimée confirme l'une des classes de l'émotion déjà apprise, ou si elle constitue une nouvelle classe (non apprise auparavant) qui nécessite une nouvelle action verbale et/ou nonverbale à synthétiser.

La deuxième partie de cette thèse porte sur les traits de la personnalité, qui jouent un rôle important dans l'interaction sociale des humains. Différentes recherches ont étudié l'effet à long terme du trait de la personnalité extraversion-introversion sur le comportement multimodal généré de l'homme. Çela le rend fiable pour caractériser le comportement verbal et non verbal combiné d'un homme en interaction avec un robot afin de permettre au robot d'adapter son comportement multimodal généré à la personnalité de l'homme. Cette adaptation de comportement pourrait subir soit le principe de la similarité d'attraction (c.-à-d., les individus sont plus attirés par d'autres qui ont des traits de personnalité similaires) soit le principe de la complémentarité d'attraction (c.-à-d., les individus sont plus attirés par d'autres dont les personnalités sont complémentaires à leurs propres personnalités) selon le contexte d'interaction. Dans cette thèse, nous examinons les effets de la multimodalité et de l'unimodalité du comportement généré sur l'interaction, en plus du principe de la similarité d'attraction puisqu'il considère l'effet de l'interaction initiale entre l'homme et le robot sur le développement de la relation entre eux (ex., l'amitié), ce qui le rend plus approprié pour notre contexte d'interaction. La détection du trait de la personnalité de l'homme comme étant introverti ou extraverti est basée sur une analyse psycholinguistique du discours de l'homme, sur laquelle les caractéristiques de la parole et des gestes générés par le robot sont définies.

Finalement, la troisième partie de cette thèse porte sur la synthèse de geste. La génération des gestes métaphoriques appropriés de tête et du bras ne suit pas une analyse linguistique spécifique. Il est principalement basé sur les indices prosodiques du discours de l'homme,

qui sont en corrélation forte avec l'émotion et les caractéristiques dynamiques de gestes métaphoriques. Le système proposé utilise les modèles de Markov-Cachés Couplés (CHMM) qui contiennent deux chaînes pour la modélisation des courbes caractéristiques segmentées de la parole et des gestes. Quand un signal-test de parole appartient au modèle de CHMM appris, un groupe correspondant de gestes métaphoriques adaptés de tête et du bras sera synthétisé. Une étude expérimentale (dans lequel le robot adapte le contenu émotionnel de son comportement multimodal généré au contexte de l'interaction) est menée pour examiner le contenu émotionnel des gestes métaphoriques du robot par l'homme directement. En outre, nous examinons les effets des expressions faciales et du discours émotionnel sur l'amélioration de l'expressivité du robot, en plus de la comparaison entre les effets de l'interaction multimodale et de l'interaction qui utilise moins d'indices affectifs sur l'homme.

En général, la recherche sur la compréhension du profil de l'homme et la génération du comportement adapté du robot ouvre la porte à d'autres sujets qui nécessitent plus d'études élaborées. Ces sujets comprennent, mais sans s'y limiter : le développement d'une architecture cognitive qui peut simuler les fonctionnalités des zones du cerveau qui permettent de comprendre et de générer la parole et les actions physiques de façon appropriée au contexte d'interaction, qui constitue une orientation future de la recherche.

TO MY FAMILY

MY LAST DEFENSE LINE

Acknowledgments

This thesis is the outcome of 4 years of hard research work as a Ph.D. candidate in the Robotics and Computer Vision laboratory at ENSTA ParisTech. During this period, I have found all the possible cooperation from each administrative and academic member of the laboratory, in addition to the friendly environment that covers all the relationships.

In the beginning, I would like to express all my deep and sincere gratitude to my supervisor **Adriana Tapus** for all the support that she offered me. She was always the nice person that encourages me to keep pushing forward despite the different difficulties and challenges that I faced, which helped me a lot to keep my passion for achieving this work in the best possible way. Similarly, I would like to acknowledge the Chaire d'Excellence program (Human-Robot Interaction for Assistive Applications) of the French National Research Agency (ANR) for supporting my work during the last years.

I would like to thank all my colleagues in the laboratory and those far colleagues whom I have met in the different scientific events that I attended all over the world one by one, for the interesting discussions we had together and the exchange of ideas that inspired me with new points in my research.

My close friends outside the laboratory have also shown all their support and encouragement for the work I am doing, especially those who have participated in the experiments that I conducted during my Ph.D. research.

Last and before all, I would like to express my sincere gratitude for my family members who gave me love and hope to succeed, especially my sisters **Asia** and **Salma**, my mother **Fawzeya**, and my father **Aly Kalfet**, whose sacrifices allowed me to reach my current position today.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human-Robot Interaction (HRI) is the field of research committed to developing and evaluating intelligent robotic systems in order to be used by humans in the daily life. This interaction, by definition, necessitates an active communication between human and robot so as to create a successful social relationship. The social communication between human and robot includes the cognitive, emotional, and personality aspects of interaction. The cognitive aspects of interaction denote the high-level functions that a robot should use for reasoning, judgment, and decision-making. This combines different research fields together, such as: linguistic analysis, cognitive psychology, artificial intelligence, and neurobiology. On the other hand, the emotional aspects of interaction refer to the robot's perception of the emotion of the interacting human or the nearby humans in the robot's environment. This requires highly sophisticated cognitive learning functions and architectures that can make the robot understand incrementally the meaning of emotion so as to generate an appropriate multimodal action corresponding to the context of interaction. Meanwhile, the personality aspects of interaction refer to the robot's perception of the interacting human's extraversion-introversion personality trait, which influences both the generated speech and gestures of human. This requires a highly sophisticated analysis for the verbal and/or non-verbal behavior of human to detect his/her extraversion-introversion level in order to make the robot generate an adapted multimodal behavior to human's personality, which could enhance the interaction between both of them.

On the other hand, the applications of human-robot interaction in human's life are increasing progressively. Robots are used nowadays in a wide variety of domains, such as: (1) medical (e.g., rehabilitation, physical therapy, surgery, and autism [Tapus et al., 2012a,b; Peca et al., 2012; Aly and Tapus, 2010]), (2) education, (3) home assistance, (4) entertainment-gaming, and (5) elderly assisted living. This increasing impact of robots in human's life requires a corresponding high-level robot's functionality in order to make the robot able to behave in a proper manner.

In this thesis, we focus on creating a basis for a long-term interactive human-robot relationship based on generating a customized multimodal robot's behavior to human's profile, which considers both the personality and emotion of human as being determinant factors for the synthesized robot's behavior.

1.1 Motivation for considering human's profile in human-robot interaction

The importance of considering emotion as a determinant factor in human-robot interaction is the fuzzy nature of emotion classes, which may have imprecise criteria of membership. This could impose a problem when designing a human-robot interaction system based on emotion detection using the traditional recognition algorithms, which may lead the robot to generate an inappropriate behavior to the context of interaction. Therefore, in this thesis, we propose an *online* fuzzy-based algorithm for detecting emotion. Besides, it precises whether a new detected emotion belongs to one of the previously learnt clusters so as to get attributed to the corresponding multimodal behavior to the winner cluster, or it constitutes a new cluster that requires a new appropriate multimodal behavior to be synthesized, as discussed in Chapter (2).

On the other hand, the long term effect of personality on the verbal and nonverbal behavior of human, makes it reliable for being considered in human-robot interaction. The adaptation of the generated multimodal robot's behavior to the extraversion-introversion personality trait of human, can increase the attraction between human and robot so as to enhance

the interaction between them. Therefore, in this thesis, we examine and validate the similarity attraction principle (i.e., individuals are more attracted by others who have similar personality traits) within a human-robot interaction context. This process of interaction integrates different subsystems that allow the robot to generate an adapted synchronized multimodal behavior to human's personality, including: a psycholinguistic-based system for detecting personality traits, and a system for generating adaptive gestures (Chapters 3 and 4).

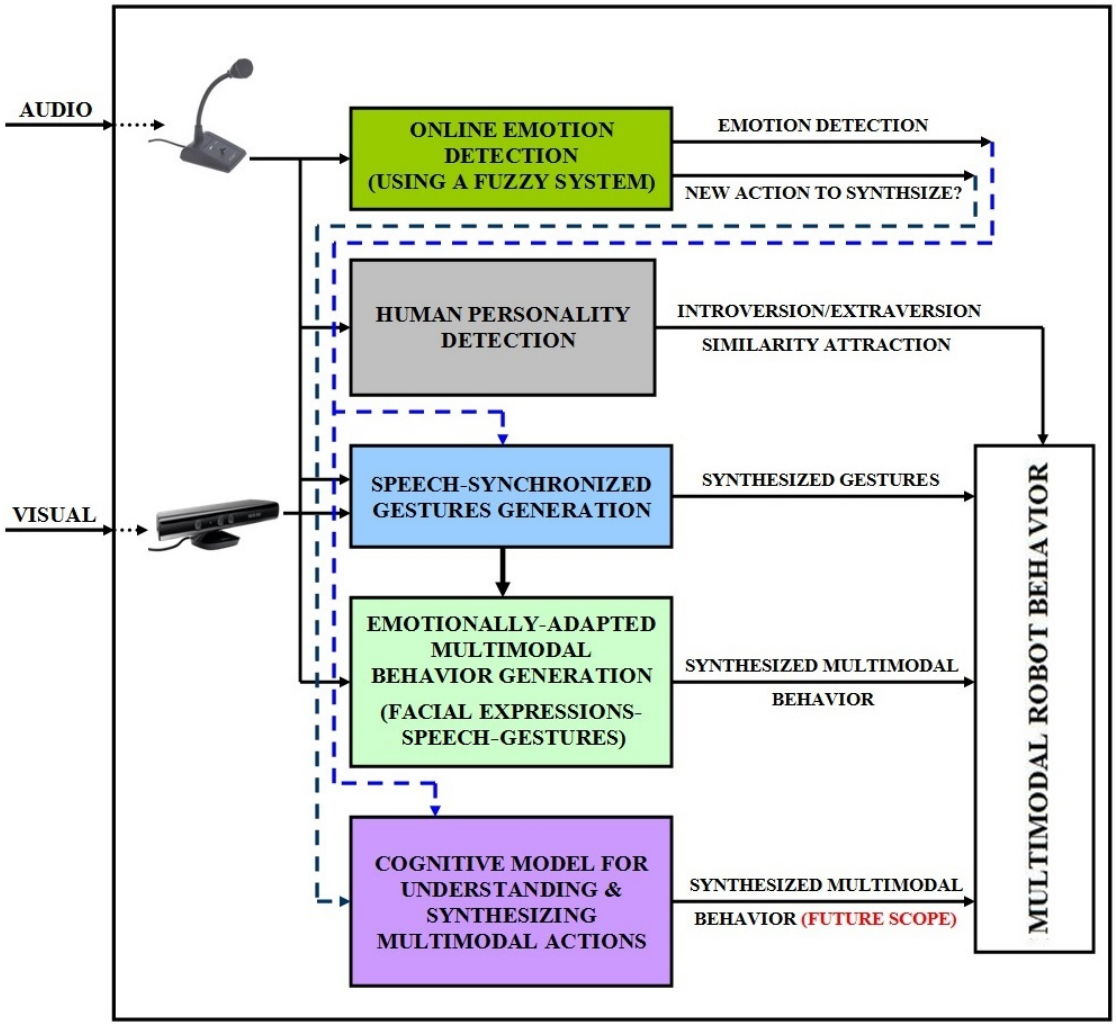


Figure 1-1 – Overview of the system architecture for generating an adapted multimodal robot's behavior [Aly and Tapus, 2014]

Figure (1-1) illustrates the structure of the proposed system in the thesis for generating an adapted multimodal robot's behavior, where the speech captured input helps in detecting

both the emotion and personality dimension of the interacting human. Besides, it intervenes (in parallel with the captured gesture input to the system) in generating a speech-synchronized multimodal behavior that combines gestures, speech, and facial expressions (Chapter 5). On the other hand, the cognitive model phase for understanding and generating multimodal actions based on a decision from the emotion detection fuzzy system, constitutes a future research scope for this thesis, as discussed in Appendix (C).

1.2 Robot testbeds

In this section, we introduce the humanoid robots used in the conducted experimental studies in the thesis:

1.2.1 NAO Robot

The humanoid NAO robot (Figure 1-2) is developed by Aldebaran Robotics[1]. NAO is a 25 degrees of freedom robot equipped with eight full-color RGB eye leds, two cameras, an inertial sensor, a sonar sensor, and many other sensors that allow it to perceive its surrounding environment with high precision and stability. This robot is employed in the studies of Chapters (2 and 3).

Figure 1-2 – NAO robot

[1] http://www.aldebaran-robotics.com/

1.2.2 ALICE Robot

The female humanoid ALICE robot (Figure 1-3) is developed by Hanson Robotics[2]. ALICE-R50 robot has a full-motion body and an expressive face, with a total of 36 degrees of freedom. The robot is equipped with two cameras and an array of sensors, including an accelerometer sensor, a torque sensor, a series of touch sensors, in addition to many other different sensors that allow it to precisely perceive its surroundings. The face of the robot composed of synthetic skin, is its main speciality. It can create a full range of credible facial expressions in different emotional states (Section 5.2.3). This robot is employed in the study of Chapter (5).

Figure 1-3 – ALICE robot

The rest of the thesis is structured in 5 chapters as following:

Chapter (2): describes an online fuzzy-based approach for detecting human's emotion.

Chapter (3): discusses synthesizing an adapted multimodal robot's behavior to human's personality.

Chapter (4): illustrates a prosody-based system for synthesizing adaptive head-arm metaphoric gestures.

[2]http://www.hansonrobotics.com/

Chapter (5): discusses synthesizing an emotionally-adapted multimodal robot's behavior within a narrative human-robot interaction.

Chapter (6): presents the overall conclusion of the conducted studies in the thesis.

Chapter 2

An Online Fuzzy-Based Approach for Human's Emotion Detection

A social intelligent robot needs to be able to understand human's emotion so as to behave in a proper manner. In this chapter, we focus on developing an online incremental learning system of emotion using Takagi-Sugeno (TS) fuzzy model. The proposed system calculates the membership values of the new data to the existing clusters, which makes it an appropriate choice for modeling ambiguous data, especially in case of any overlapping clusters (e.g., emotion clusters). The main objective of this system is to detect whether an observed emotion needs a new corresponding multimodal behavior to be synthesized in case it constitutes a new emotion cluster not learnt before, or it can be attributed to the corresponding multimodal behavior to an existing cluster, to which the new observed emotion belongs. The online evolving fuzzy rules of the TS model are updated incrementally whether by modifying the previously learnt rules, or by adding new rules according to the cluster centers of the new data, in which each cluster center represents a rule in the TS model. Consequently, the total number of rules in the TS model could be increased in case a new cluster center is accepted. Meanwhile, a previously learnt rule could be replaced or modified according to the descriptive potential of the new data. The obtained results show the effectiveness of the proposed methodology.

2.1 Introduction

The fast developing human-robot interaction (HRI) applications require the robot to be capable of behaving appropriately in different and varying situations as humans do. This objective necessitates the robot to have high level cognitive functions so as to make it able to detect the emotion of the interacting human in order to generate a corresponding action.

Human's emotional states detection has been a rich research topic during the last decade. Traditional approaches are based on constructing a finite database with a specific number of classes, and on performing a batch (offline) learning of the constructed database. However, the associated problems with the batch learning show the importance of processing data online for the following reasons: (1) avoiding storage problems associated with huge databases, and (2) input data comes as a continuous stream of unlimited length, which creates a big difficulty in front of applying the batch learning algorithms. The absence of online learning methods can make the robot unable to cope with different situations in an appropriate way due to an error in classifying a new emotion as being one of the previously learnt emotions, while its content constitutes a new emotional state category.

Many approaches are present in the literature for the detection of human's affective states from voice signal. The significance of prosody in conveying emotions is illustrated in Cahn [1990a] and Murray and Arnott [1993]. The authors discussed a comparative study about the variation of some relevant parameters (e.g., pitch and voice quality) in case of different emotional states. Moreover, Cahn [1990a] explained the emotionally driven changes in voice signal's features under physiological effects in order to understand how the vocal (i.e., tonal) features accompanying emotions could differ. Roy and Pentland [1996] illustrated a spoken affect analysis system that can detect speaker's approval versus speaker's disapproval in child-directed speech. Similarly, Slaney and McRoberts [1998] proposed a system that can recognize prohibition, praise, and attentional bids in infant-directed speech. Breazeal and Aryananda [2002] investigated a more direct scope for affective intent recognition in robotics. They extracted some vocal characteristics (i.e., pitch and energy) and discussed how they can change the total recognition score of the affective intent in robot-

directed speech. A framework for human's emotion recognition from voice through gender differentiation was also described in Vogt and Andre [2006]. Generally, the results of the offline recognition of emotions in terms of the above mentioned vocal characteristics, are reasonable. On the other hand, emotion-based applications became more and more important. In computer-based applications, the interacting system needs to recognize human's emotion in order to generate an adapted behavior so as to maintain the maximum engagement with human [Voeffra, 2011; Clavel et al., 2012]. Similarly, emotion-based applications appear in different areas that engage both human and robot, like: entertainment, education, and general services [Pierre-Yves, 2003; Jones and Deeming, 2008].

On the other hand, the importance of using fuzzy logic in modeling complex systems has been increased gradually in the last decade. It imitates the human logic using a descriptive and imprecise language in order to cope with input data. Zadeh [1965, 1973] put the first theory of fuzzy sets after observing that the traditional mathematical definition of object classes in real world is neither sufficient nor precise, because these classes may have imprecise criteria of membership. This observation remains valid for emotion classes, so that the emotion class "anger" may have clear membership criteria in terms of its vocal (i.e., tonal) characteristics with respect to the emotion class "sadness". However, it can have ambiguous membership criteria when compared to the emotion class "happiness" because of the vocal characteristics' similarity of the two emotional states. One of the main reasons behind this ambiguity is that people show different amounts of spoken affect according to their personal and cultural characteristics [Peter and Beale, 2008]. This validates the necessity of modeling emotions using fuzzy sets and linguistic *if-then* rules in order to illustrate the existing fuzziness between these sets. Fuzzy inference is the process of mapping an input to a corresponding output using fuzzy logic, which formulates a basis for taking decisions. The literature of fuzzy inference systems reveals two major inference models: Mamdani and Assilian [1975] and Takagi and Sugeno [1985]. Mamdani and Assilian [1975] stated the first fuzzy inference system designed for controlling a boiler and a steam engine using a group of linguistic control rules stated by experienced human operators. Meanwhile, Sugeno [1985] and Takagi and Sugeno [1985] proposed another fuzzy inference system known as TS fuzzy model, which can generate fuzzy rules from

a given input-output dataset. Clearly, TS fuzzy model is the model adopted in this study, because we have an initial database of emotion labeled states constituting the input-output data necessary for defining the initial TS model. The relationship between these emotional states is represented by fuzzy sets. When new data arrives, whether a new TS model is constructed corresponding to a newly created cluster, or one of the existing TS models is updated according to the cluster, to which the new data is attributed.

On the way for an online recognition system of human's emotional states, clustering algorithms have proven their importance [Bezdek, 1981; Vapnik, 1998]. Clustering implicates gathering data vectors based on their similarity. It generates specific data points "cluster centers" that construct the initial TS fuzzy rules indicated above. K-means algorithm defines the membership of each data vector as being related to one cluster only, in addition to not belonging to the rest of the clusters. Fuzzy C-means algorithm, which was first proposed by Dunn [1973], then improved by Bezdek [1981], is an extension of the K-means algorithm that considers the fuzziness existing within a dataset. Consequently, it indicates the membership degrees of data vectors to all the existing clusters. However, both the dataset and number of clusters are required to be defined a priori, which makes it not applicable for our online recognition approach. Gustafsson and Kessel [1979] developed the classical Fuzzy C-means algorithm using an adaptive distance norm in order to define clusters of different geometrical shapes within a dataset. However, similarly to the Fuzzy C-means algorithm, the number of clusters is required to be defined a priori. Furthermore, Gath and Geva [1989] described an unsupervised extension of the algorithm illustrated in Gustafsson and Kessel [1979] (which takes both the density and size of clusters into account), so that a priori knowledge concerning the clusters' number is no longer required. However, this methodology suffers from other problems, such as: (1) the algorithm can get easily stuck to the local minima with increasing complexity, and (2) it is difficult to understand the linguistic terms defined through the linear combination of input variables.

Other algorithms were proposed in order to overcome the drawbacks of the previously mentioned clustering algorithms. For example, the mountain clustering algorithm [Yager and Filev, 1992, 1993] tries to calculate cluster centers using a density measure (mountain

function) of a grid over the data space, in which cluster centers are the points with the highest density values. However, even if this algorithm is relatively efficient, its computational load increases exponentially with the dimensionality of the problem. The subtractive clustering [Chiu, 1994] solved this problem by considering data points as possible candidates for cluster centers, instead of constructing a grid each time when calculating a cluster center, as in the mountain clustering. In this work, we chosed to use the subtractive clustering algorithm in order to identify the parameters of the TS fuzzy model [Takagi and Sugeno, 1985; Chiu, 1994].

The rest of the chapter is structured as following: Section (2.2) illustrates a general overview for the basic and complex emotions, Section (2.3) discusses the offline detection of human's emotional states, Sections (2.4) and (2.5) overview the subtractive clustering and Takagi-Sugeno fuzzy model, Section (2.6) describes the online updating of Takagi-Sugeno fuzzy model, Section (2.7) provides a description of the results, and finally Section (2.8) concludes the chapter.

2.2 Basic and complex emotions

Emotion is one of the most controversial issues in human-human interaction nowadays, in terms of the best way to conceptualize and interpret its role in life. It seems to be centrally involved in determining the behavioral reaction to social environmental and internal events of major significance for human [Izard, 1971; Plutchik, 1991]. One of the main difficulties behind studying the objective of emotion is that the internal experience of emotion is highly personal and is dependent on the surrounding environment circumstances. Besides, many emotions may be experienced simultaneously [Plutchik, 1991].

Different emotion theories identified relatively small sets of fundamental or basic emotions, which are meant to be fixed and universal to human (i.e., they can not be broken down into smaller parts). However, there is a deep opinion divergence regarding the number of basic emotions. Ekman [1972]; Ekman et al. [1982] stated a group of 6 fundamental emotions (i.e., anger, happiness, surprise, disgust, sadness, and fear) after studying cross-cultural

facial expressions, collected from a lot of media pictures for individuals from different countries. However, Ekman in his theory had not resolved the problem discussed in the research of Izard [1971], which is the fact that it is not possible, or at least not easy, to unify basic universal facial expressions through processing media pictures only, because there are a lot of populations who have no access to media, like some populations in Africa. Consequently, there is no considerable database for their facial expressions to study. Thereafter, Izard [1977] devised a list of 10 primary emotions (i.e., anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise), each one has its own neurobiological basis and pattern of expression (usually denoted by facial expressions), and each emotion is experienced uniquely. Tomkins [1984] proposed a biologically-based group of pan-cultural 9 primary emotions (i.e., anger, interest, contempt, disgust, distress, fear, joy, shame, and surprise). More theories exist in the literature of emotion modeling, similarly to the previously stated theories. However, they do not consider the evolutionary and combinatory nature of emotion, which may lead to a new advanced category of complex emotions that could be considered as mixtures of primary emotions based on cultural or idiosyncratic aspects.

Plutchik proposed an integrative theory based on evolutionary principles [Plutchik, 1991]. He created a three-dimensional (i.e., intensity, similarity, and polarity) circumplex wheel of emotions that illustrates different compelling and nuanced emotions based on a psychological-biological research study, as indicated in Figure (2-1). The eight sectors of the wheel indicate that there are 8 primary emotions (i.e., anger, fear, disgust, trust, sadness, joy, surprise, and anticipation), arranged in four opposite pairs (i.e., different polarity; e.g., joy versus sadness). The circles represent emotions of similar intensity; the smaller circle contains the emotions of highest intensity in each branch, while the second circle contains extensions of the first circle's emotions, but in lighter intensity, and so on. The blank spaces represent the primary dyads, which are mixtures of two adjacent primary emotions. However, the secondary dyads are mixture of two non-adjacent primary emotions with one primary emotion in-between (e.g., $anger + joy = pride$, or $fear + sadness = desperation$). Meanwhile, tertiary dyads are mixtures of two non-adjacent primary emotions with two primary emotions in-between (e.g., $fear + disgust = shame$, or $anticipation + fear = anxiety$). Plutchik model is, therefore, the most appropriate model for our research.

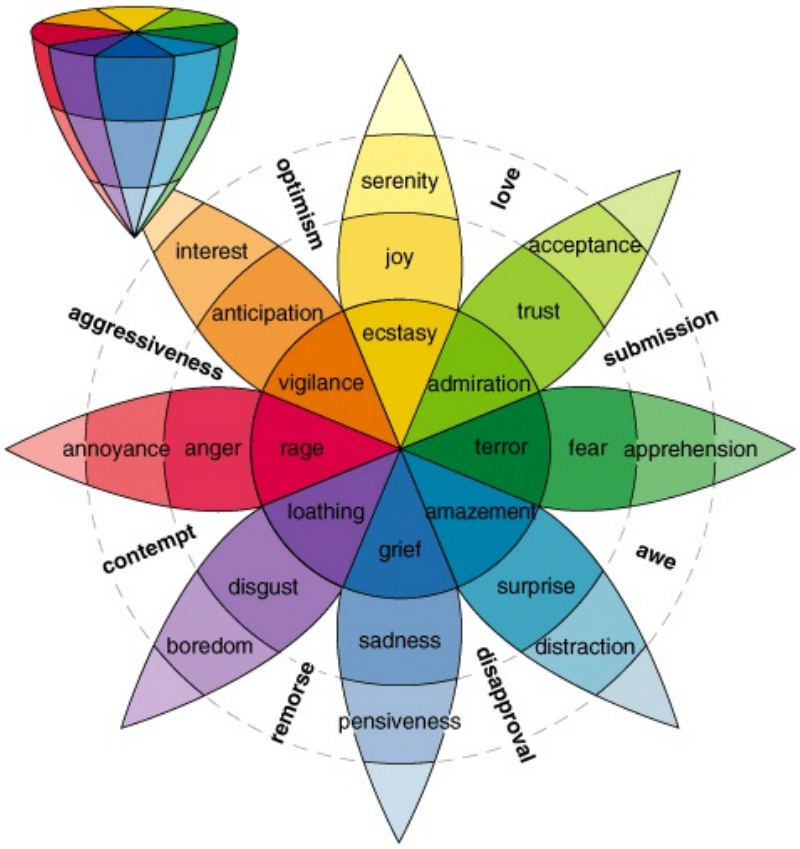


Figure 2-1 – Plutchik's primary and mixture emotions presented in a 2D wheel, and in a 3D cone [Plutchik, 1991]

2.3 Offline detection of emotional states

In this chapter, we investigated the performance of the offline classification system using the Support Vector Machine (SVM) algorithm [Cortes and Vapnik, 1995], with 15 primary and complex emotions. Afterwards, we created a fuzzy classification system and we trained it offline on 6 primary emotions, in addition to the neutral emotion (i.e., anger, disgust, happiness, sadness, surprise, fear, and neutral). However, the online test phase of the fuzzy model contained 5 complex emotions (i.e., anxiety, shame, desperation, pride, and contempt), in addition to 3 primary emotions (i.e., interest, elation, and boredom).

Three databases (including more than 1000 voice sample) have been employed in training and testing the classification system. These databases are: (1) German emotional speech database (GES) [Burkhardt et al., 2005], (2) Geneva vocal emotion expression stimulus

set (GVEESS) [Banse and Scherer, 1996] [1], and (3) Spanish emotional speech database (SES) [Montero et al., 1998]. [2] An important remark about the emotion classes of the total database is that they do not have all the same intensity, in addition to the existing emotion extension in two cases: boredom-disgust, and elation-happiness. This is due to the encountered difficulty to obtain well known databases with specific emotion categories that exactly match Plutchik model's emotion categories.

Relevant vocal features (i.e., pitch and energy curves [3]) have been calculated for all the samples of the databases [Breazeal and Aryananda, 2002] in order to find out their possible effects on characterizing emotional states. The emotional state detection system, normally, includes three different subprocesses: speech signal processing (Section 2.3.1), features extraction (Section 2.3.2), and classification (Section 2.3.3), as indicated in Figure (2-2).

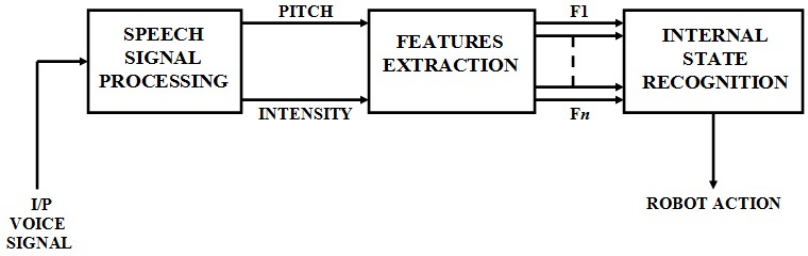


Figure 2-2 – Emotional state detection system

2.3.1 Speech Signal Processing

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 temporal and spectral distribution of energy. Instead, another vocal (i.e., tonal) characteristic, which is the fundamental frequency $F0$, is calculated 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

[1] The stimulus set used is based on research conducted by Klaus Scherer, Harald Wallbott, Rainer Banse and Heiner Ellgring. Detailed information on the production of the stimuli can be found in [Banse and Scherer, 1996].

[2] The SES database is a property of Universidad Politecnica de Madrid, Departamento de Ingenieria Electronica, Grupo de Tecnologia del Habla, Madrid (Spain).

[3] We use the terms energy and intensity interchangeably between the chapters of the thesis.

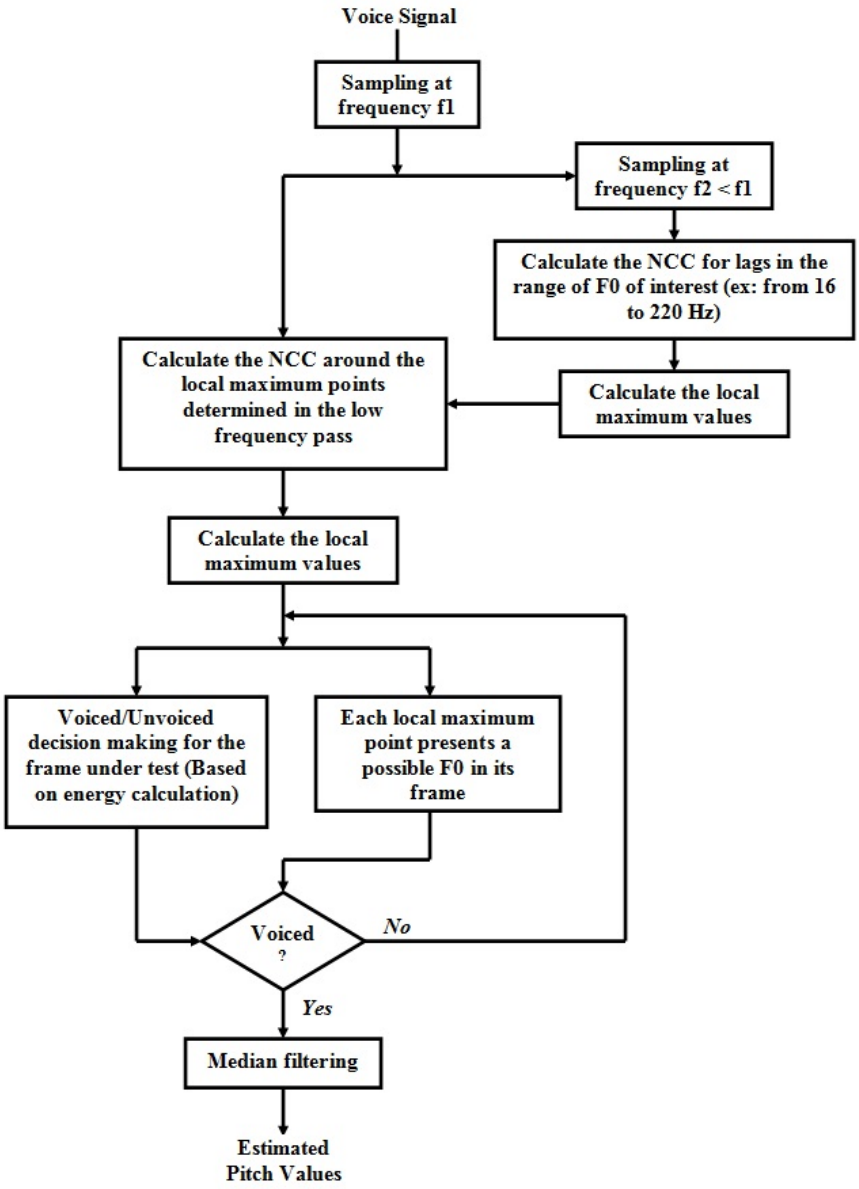


Figure 2-3 – Pitch tracking

processing. The signal conditioning process tries to clear away interfering signal components, such as any unnecessary noise using low pass filtering, which removes any loss of periodicity in the voiced signal's spectrum at high frequencies, and using high pass filtering when there are DC or very low frequency components in the signal. The candidate periods estimation step tries to estimate the candidate voiced periods from which the fundamental frequency $F0$ could be calculated. Talkin [1995] developed the traditional Normalized Cross Correlation (NCC) method [Sondhi, 1968; Rabiner et al., 1977] in order to estimate

reliably the voicing periods and the fundamental frequency $F0$ by considering all candidates simultaneously in a large temporal context in order to avoid the variation of the glottal excitation periods through the signal. This methodology uses a two pass NCC calculation for searching the fundamental frequency $F0$, which reduces the overall computational load with respect to the traditional NCC methodology. Finally, the post processing step uses median filtering in order to refine the calculated fundamental frequency $F0$ and ignore isolated outliers, as indicated in Figure (2-3). On the other hand, voice signal's energy could be calculated from squaring the amplitude of signal's points.

2.3.2 Features Extraction

Rong et al. [2008] presented a detailed study concerning the common vocal (i.e., tonal) characteristics used in the literature of emotion recognition, and their significance. After testing different tonal characteristics in the offline classification phase, we found that the most important characteristics are: pitch and energy, upon which the recognition score highly depends. Meanwhile, other characteristics (e.g., duration and rhythm) did not have the same significant effect on the recognition score. Relevant statistical measures of signal's pitch and energy were calculated to create the characteristic vectors used in constructing the database. The final features used in this work are: (1) pitch mean, (2) pitch variance, (3) pitch maximum, (4) pitch minimum, (5) pitch range, (6) pitch mean derivative, (7) energy mean, (8) energy variance, (9) energy maximum, and (10) energy range.

2.3.3 Classification

Voice samples were classified using the SVM algorithm with a quadratic kernel function [Platt, 1998; Cristianini and Shawe-Taylor, 2000], and the results were cross validated. Table (2.1) indicates the obtained recognition scores of 15 different emotions. The mean values of the recognition scores indicated in Table (2.1), reflect the high precision of our classification system with respect to similar systems discussed in the literature. In Burkhardt et al. [2005], the mean value of the emotion recognition scores was 86.1% [Aly and Tapus, 2011c]. Meanwhile, in Banse and Scherer [1996], the mean value of the scores was 60%. However, in Montero et al. [1998], the mean value of the obtained scores was 85.9%.

The calculated recognition scores of emotions depend mainly on the individuals performing the emotions, and on the amount of spoken affect they show. This may lead to a problem in real human-robot interaction scenarios if the expressed emotion to the robot is different (in terms of its tonal features) from the trained emotion in the database. Consequently, two scenarios may exist: (1) if the expressed emotion is intended to belong to one of the existing emotion classes in the database, it is probable that the robot misclassifies it. This depends totally on the performance of the recognition system, and (2) if the expressed emotion does not belong to any of the existing emotion classes in the database, and the robot attributes it to the nearest existing emotion class (instead of constituting a new emotion class), it may lead the robot to behave in an inappropriate manner. Therefore, in order to avoid any improper robot's behavior, it is important for the robot to understand whether the online expressed emotion constitutes a new emotional state class or not. This allows the robot to perform a neutral action different from the corresponding prescribed actions to the learnt emotions so as to not make the performed action seems to be out of context to the interacting human (developing autonomously an appropriate multimodal robot's affective behavior is slightly discussed in Appendix (C) as a future research scope for this thesis).

Emotion	GES	GVEESS	SES	All 3 DB mixed
Anger	80.8%	88.7%	79.8%	81.7%
Boredom	85.4%	87.1%	-	90.1%
Disgust	92.1%	91.7%	-	93.5%
Anxiety	87.3%	86.5%	-	87.5%
Happiness	86.9%	88.5%	75.1%	86.1%
Neutral	83.7%	-	89.5%	87.8%
Sadness	86.9%	90.1%	94.1%	85.7%
Surprise	-	-	95.7%	96.3%
Interest	-	89.3%	-	90.4%
Shame	-	90.7%	-	91.9%
Contempt	-	91.3%	-	90.6%
Desperation	-	87.7%	-	89.2%
Elation	-	89.9%	-	87.5%
Pride	-	86.9%	-	87.3%
Fear	-	85.7%	-	89.7%
Mean Value	86.2%	88.8%	86.8%	89%

Table 2.1 – Recognition scores of different emotional states. Empty spaces represent the not included emotions in these databases.

2.4 Subtractive clustering

Subtractive clustering [Chiu, 1994] is an algorithm used for calculating cluster centers within a dataset. It uses data points as possible candidates for cluster centers, and then it calculates a potential function for each proposed cluster center, which indicates to what extent the proposed cluster center is affected by the surrounding points in the dataset. Suppose a cluster composed of k normalized data points $\{x_1, x_2, ..., x_k\}$ in an M-dimensional space, where each data point has a potential P that could be represented as following in Equation (2.1):

$$P_d = \sum_{u=1}^{k} e^{-\frac{4}{r^2}\|x_d - x_u\|^2}; \qquad d \in \{1 \cdots k\} \tag{2.1}$$

where r is the neighborhood radius that is fixed to 0.3, at which the calculation of cluster centers is optimally precise. After choosing the first cluster center (which is the data point with the highest potential value), the potential of the other remaining data points will be recalculated with respect to it.

Assume x_n^* is the location of the n^{th} cluster center of potential P_n^*, consequently the potential of each remaining data point could be formulated as following (Equation 2.2, where r_b is a positive constant):

$$P_d \Leftarrow P_d - \underbrace{P_n^* e^{-\frac{4}{r_b^2}\|x_d - x_n^*\|^2}}_{X} \tag{2.2}$$

From the previous equation, it is clear that the potential of each remaining data point is subtracted by the amount X, which is a function of the distance between the point and the last defined cluster center. Consequently, a data point near to the last defined cluster center will have a decreased potential, so that it will be excluded from the selection of the next cluster center. In order to avoid having close cluster centers, the value of r_b should be chosen greater than the value of the neighborhood radius r (r_b=1.5r) [Chiu, 1994]. After calculating the reduced potential of all data points with respect to the last defined cluster center according to Equation (2.2), the next cluster center is chosen as the new highest

potential value. This process is repeated until a sufficient number of centers is attained.

Chiu [1994] proposed a criterion for accepting and rejecting cluster centers in order to define the final sufficient number of clusters. This criterion defines two limiting conditions: lower ($\underline{\varepsilon}P_1^*$) and upper ($\overline{\varepsilon}P_1^*$) boundaries (where $\overline{\varepsilon}$ and $\underline{\varepsilon}$ are small threshold fractions). A data point is selected to be a new cluster center if its potential is higher than the upper threshold, and is rejected when its potential value is lower than the lower threshold. If the potential of the data point is between the upper and lower thresholds, a new decisive rule is used for accepting new cluster centers (Equation 2.3):

$$\frac{d_{min}}{r} + \frac{P_n^*}{P_1^*} \geq 1 \tag{2.3}$$

where d_{min} is the shortest distance between x_n^* and the locations of all the previously calculated cluster centers. Otherwise, the data point is rejected.

According to Chiu [1994], the upper threshold ($\overline{\varepsilon}$) is fixed to 0.5, while the lower threshold ($\underline{\varepsilon}$) is fixed to 0.15. This approach is used for calculating the antecedent parameters of the fuzzy model. It depends on the fact that each cluster center represents a characteristic fuzzy rule for the system.

2.5 Takagi-Sugeno (TS) fuzzy model

Takagi-Sugeno (TS) fuzzy model employs fuzzy rules, which are linguistic statements ($if-then$), involving fuzzy logic, fuzzy sets, and fuzzy inference. The fuzziness in the input sets is characterized by the input membership functions, which could have varying shapes (triangular, Gaussian, etc.) according to the nature of the modeled process.

Considering a set of n cluster centers $\{x_1^*, x_2^*, ..., x_n^*\}$ produced from clustering the input-output data space; each vector x_i^* is decomposed into two component vectors y_i^* and z_i^*, which contain the cluster center's coordinates in the input and output spaces in order (i.e., the number of the input and output membership functions is determined by the number of cluster centers).

Suppose that each cluster center x_i^* is a fuzzy rule, therefore for an input vector y= [y_1, y_2, ..., y_m], the firing degree of the input vector's component y_j to the input membership function corresponding to the j^{th} input component and the i^{th} fuzzy rule y_{ji}^* is defined as following (Equation 2.4) [Angelov, 2002]:

$$\mu_{ji} = e^{(-\frac{4}{r^2}\|y_j - y_{ji}^*\|^2)}; \quad i \in \{1 \cdots n\}, \quad j \in \{1 \cdots m\} \tag{2.4}$$

Consequently, the total degree of membership of rule i with respect to the whole input vector could be defined as following (Equation 2.5):

$$\tau_i = \mu_{1i}(y_1) \times \mu_{2i}(y_2) \times \cdots \times \mu_{mi}(y_m) = \prod_{j=1}^{m} \mu_{ji}(y_j) \tag{2.5}$$

The previous model could be reformulated in terms of linguistic *if-then* fuzzy rule as following (Equation 2.6):

$$\begin{aligned} &\text{If } y_1 \text{ is } y_{1i}^* \text{ and } \cdots\cdots \text{ and } y_m \text{ is } y_{mi}^* \\ &\text{Then} \quad z_i^* = b_{0i} + b_{1i}y_1 + \cdots + b_{mi}y_m \end{aligned} \tag{2.6}$$

where z_i^* is the corresponding linear output membership function to rule i.

The input membership functions represent generally a linguistic description of the input vector (e.g., small, big, etc.). Therefore, the first antecedent part of the rule (y_1 is $y_{1i}^* \cdots$) represents the membership level of the input y_1 to the function y_{1i}^*. The output vector z could be represented in terms of the weighted average of rules contributions as following (Equation 2.7):

$$z = \sum_{i=1}^{n} \frac{\tau_i z_i^*}{\sum_{l=1}^{n} \tau_l} = \sum_{i=1}^{n} \gamma_i z_i^* \tag{2.7}$$

The learning parameters of the consequent part of the rule could be estimated by the recursive least squares approach. Suppose $\lambda_i = [b_{0i}, b_{1i}, \cdots, b_{mi}]$, $Y = [1, y_1, \cdots, y_m]^T$, so that

the previous equation could be reformulated in terms of all the fuzzy rules as following (Equation 2.8):

$$z = \chi\varphi \tag{2.8}$$

where $$\chi = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix}, \qquad \varphi = [\gamma_1 Y, \gamma_2 Y, \cdots, \gamma_n Y]$$

In our context, for an existing emotional state cluster, the given set of input-output data is used to define a cost function, from which the parameters set χ could be calculated by minimizing that function (Equation 2.9, where k is the number of data points within a cluster):

$$J = \sum_{d=1}^{k} (z_d - \chi\varphi_d)^2 \tag{2.9}$$

Equation (2.9) could be reformulated as following (Equation 2.10, where the matrices Z, η are functions in z_d and φ_d):

$$J = (Z - \chi\eta)^T (Z - \chi\eta) \tag{2.10}$$

The least square estimation of χ, could be finally defined as following (Equation 2.11):

$$\hat{\chi} = (\eta\eta^T)^{-1}\eta Z \tag{2.11}$$

A typical fuzzy modeling of a human's emotional state is illustrated in Figure (2-4), in which each vocal feature is mapped to a corresponding group of input membership functions equal to the number of rules. The output of the model is represented by the value of z calculated in Equation (2.7). When the vocal features of a test voice sample are calculated, they get evaluated through the fuzzy model of each existing emotion. The decisive criterion of the emotional state's class to which the voice sample is attributed, could be defined as

following (Equation 2.12, where α is the total number of the existing clusters):

$$Class = \arg\max_{p=1}^{\alpha}(z_p) \tag{2.12}$$

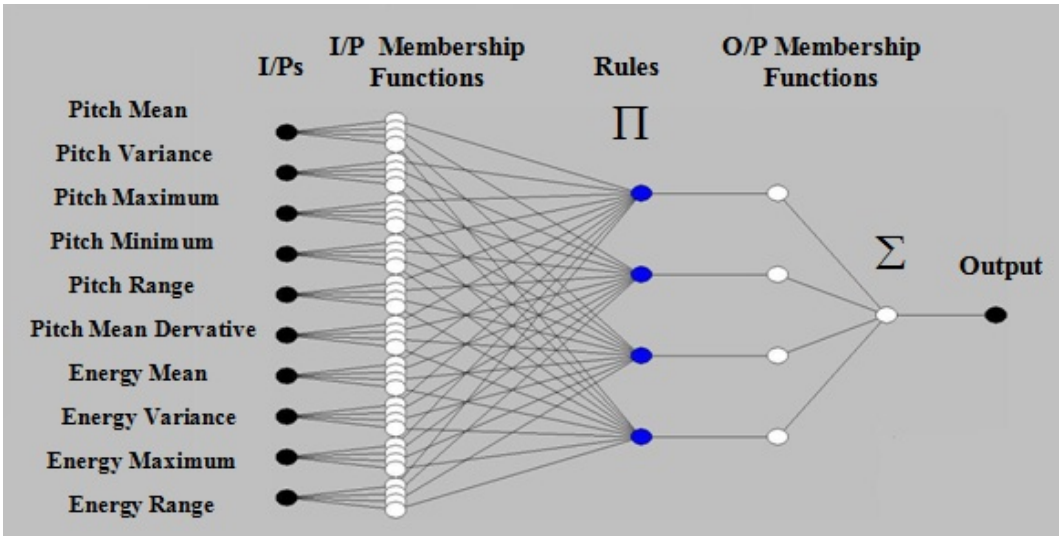


Figure 2-4 – TS fuzzy modeling of a human's emotion cluster

2.6 TS fuzzy model online updating

The online updating of the constructed TS fuzzy model is essential for continuous data streams. This requires an incremental calculation for the informative potential of the on-line incoming data [Angelov, 2002] in order to decide whether the new data confirms the contained information in one of the existing data clusters, or it constitutes a new cluster (Figure 2-5). When a new data element arrives, it gets attributed to one of the existing clusters according to Equation (2.12), which leads to one of the three scenarios below:

2.6.1 Scenario 1

A new data element is attributed with a good score to an existing emotion cluster, so that the robot implements the associated action with the winner class. Considering the emotion recognition scores shown in Table (2.1), and the possible variation in the spoken affect shown by humans in real interaction experiments, we considered this score to be $> 80\%$ in

order to assure a relatively high confidence in recognizing emotions. On the other hand, the fuzzy model of the winner class should keep updated in order to get ready for the arrival of any new element to the model (Figure 2-5). The procedures of updating the TS model are summarized in the following pseudo code (where n is the number of cluster centers):

1: **if** $(P_{NEW} > P_l^*), \forall\, l \in \{1 \cdots n\}$ and the new data point is near an old cluster center, so that the following inequality is fulfilled: $\frac{P_{NEW}}{max P^*_{l\in\{1\cdots n\}}} - \frac{d_{min}}{r} \geq 1$ **then** the new data point will replace the old rule center.
go to: Scenario 3.

2: **else if** $(P_{NEW} > P_l^*), \forall\, l \in \{1 \cdots n\}$ **then** the new point will be considered as a new cluster center x^*_{NEW}, thus a new fuzzy rule will be created.
go to: Scenario 3.

3: **else** The new data point does not possess enough descriptive potential to update the model, neither by creating a new rule, nor by replacing one of the existing rules.

4: **end if**

For the steps 1 and 2 of the pseudo code, the consequent parameters of the TS model should be estimated recursively, as indicated in Equations (2.7) to (2.11). Similarly, for all the steps 1, 2, and 3, the potential of all cluster centers needs to be calculated recursively. This is due to the fact that potential calculation measures the density level of groupings in the data space, consequently this measure will be reduced for a given cluster center in case the data space gets increased by acquiring more data elements of different patterns (Equation 2.2). Typically, the potential of a new acquired data point P_{NEW} will be increased, when other new data points of similar patterns group with it [Angelov, 2002].

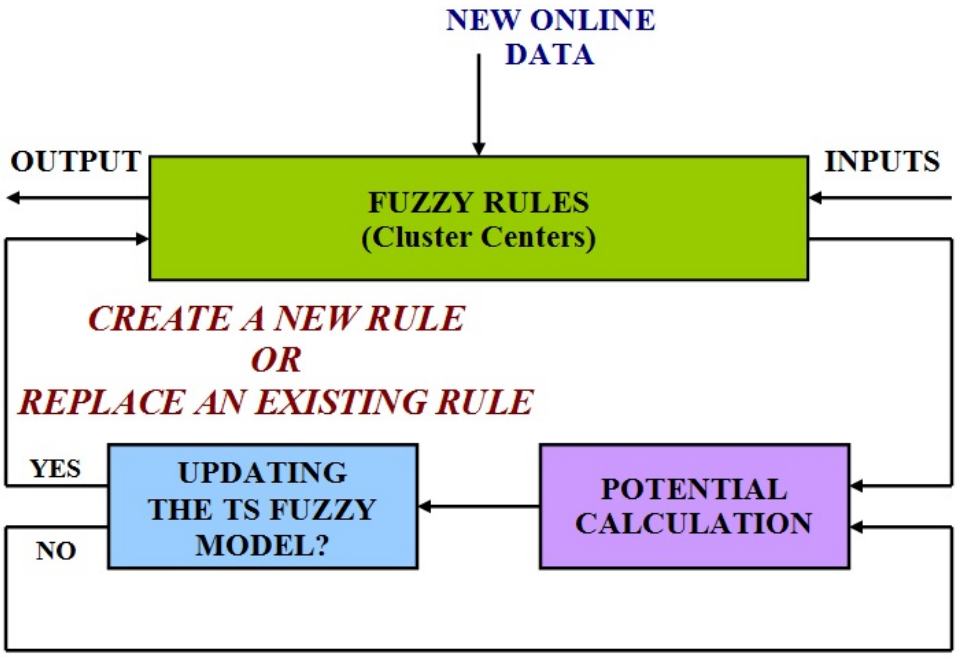


Figure 2-5 – TS fuzzy model updating whether by creating a new rule, or by replacing an existing rule.

2.6.2 Scenario 2

In case the recognition score of the existing clusters for a new data element does not reach the predefined threshold (i.e., $< 80\%$), an uncertainty factor would be considered. Consequently, the new data element will be attributed temporarily to all the existing clusters at the same time with a specific label in order to distinguish it from the normal data elements of each cluster. Afterwards, the robot will implement a prescribed neutral action (different from the normal neutral action associated with the “neutral" emotion class), until its cognitive awareness increases and gets ready to synthesize its own multimodal action according to the context as referred to earlier.

The main reason behind attributing temporarily the new data element X_{NEW} to all the existing clusters is that when the potential of this new element is recursively calculated, it gets increased gradually when other uncertain data elements get attributed, in a similar manner, to all clusters, provided that they have a similar data pattern as X_{NEW}. Meanwhile, the potential of clusters’ original centers will be reduced (Equation 2.2). Consequently, a new cluster will be created (with an associated neutral action, until the robot gets able to synthe-

size an alternative action by its own), if the potential of X_{NEW} gets greater than the potential of all the original centers in each cluster, as indicated in the following pseudo code (where α is the number of the existing clusters, and n is the number of cluster centers):

1: **if** $(P_{P_{NEW}} > P^*_{p,l}), \forall\, l \in \{1 \cdots n\},\ \ p \in \{1 \cdots \alpha\}$
then all the copies of the uncertain new data elements with similar patterns will be removed from all clusters, and only one group of them will create the new cluster.
and $\alpha := \alpha$+1
and A new TS fuzzy model will be created for the new cluster.
go to: Scenario 1.
2: **end if**

In case a new element gets attributed to a specific cluster with a confident score as in Scenario 1, the existence of temporarily uncertain data elements in this cluster will not affect the potential calculation of the new data element with respect to the original cluster centers. Therefore, they will not participate (in this case) in updating the TS fuzzy models of the clusters in which they exist, which explains the reason behind being labeled differently.

2.6.3 Scenario 3

During Scenario 2, it is possible that one of the uncertain data elements was belonging originally to one of the existing clusters, and got classified as an element of uncertain emotional content though, due to the lack of experience. This results from fact that people show emotional affect in different ways even for the same expressed emotion, which may create a problem that is the necessity to train the classifier on unlimited emotion patterns for each cluster. Consequently, it is probable that the previous learning experience of the classifier was not sufficient enough to recognize the new data element with a confident score. In order to avoid this problem, at each moment when a cluster is updated by a new element recognized with a confident score as in Scenario 1, a revision on the uncertain elements of

this cluster will be performed by re-calculating the recognition scores of the updated cluster's fuzzy model for the uncertain elements. If any uncertain element is recognized with a confident score by the fuzzy classifier of the updated cluster, this element will join the updated cluster, and its copies will be eliminated from the uncertain data spaces of all other clusters, as indicated in the following pseudo code (where ω is the number of cluster's uncertain data points, S denotes the recognition score, and k is the number of the cluster's certain data points):

1: **do** Scenario 1 (steps 1 and 2)
2: **if** $(S_{P,u} > 80\%), \forall\, u \in \{1 \cdots \omega\},\ \ p \in \{1 \cdots \alpha\}$
   then the uncertain data point $x_{p,u}$ will join the correct cluster, and will be removed from all the other clusters.
   and $k_P := k_P{+}1$
   go to: Scenario 1.
3: **end if**

2.7 Results and discussion

The fuzzy classification system was trained on 7 emotions (i.e., anger, disgust, happiness, sadness, surprise, fear, and neutral), and the results were cross validated (Table 2.2). The calculated scores are less than the previously obtained scores through the offline learning process using the SVM algorithm (Table 2.1), because the SVM algorithm deals directly with the data space, meanwhile the fuzzy classification system deals with the data space through an approximate TS model, however they remain acceptable results.

The online test database included voice samples covering simple and complex emotions from the three databases referred to earlier (Section 2.3), in addition to some other voice samples (for the same emotions), expressed by other actors in a noisy environment in our laboratory. These 8 emotions are: anxiety, shame, desperation, pride, contempt, interest, elation, and boredom. Table (2.3) illustrates the results of attributing the test clusters' data elements to the existing old clusters, upon which the system was trained on. A small part

of the test data elements was attributed with a confident score (i.e., $> 80\%$) to the existing clusters, which is unavoidable and depends totally on the patterns of the test data elements, and on the actors' performance. However, the results of classification are not totally out of context, like: the elements of the "anxiety" class that were attributed to the "fear" class, and the elements of the "elation" class that were attributed to the "happiness" class.

Emotion	Recognition Score
Anger	83.76%
Disgust	75.60%
Happiness	76.92%
Sadness	69.57%
Surprise	80.28%
Fear	77.08%
Neutral	82.14%
Mean Value	77.91%

Table 2.2 – Recognition scores of the fuzzy system's training emotions

The part of the new data attributed to the existing clusters (Table 2.3), was assigned for the validation of Scenario 1 (Section 2.6). The main encountered problem was that the new data elements attributed to the existing clusters were generally too few to update the fuzzy models of clusters easily. Unlike the elements attributed to the "fear" class, which were sufficiently descriptive to update the fuzzy model, so that two new elements satisfied the steps 1 and 2 of Scenario 1. On the other hand, the uncertain part of the new data (Table 2.3), was assigned for the validation of Scenario 2 (Section 2.6). Two new clusters were successfully constructed in case of the "anxiety" and "boredom" emotions. To the contrary, the number of elements in the other classes (i.e., shame, desperation, pride, contempt, interest, and elation), was not sufficient to fulfill Scenario 2. Therefore, the elements of these classes were considered as uncertain data elements, until more data elements of similar patterns were acquired, then Scenario 2 was re-checked.

A video showing our system working in a simple interaction experiment with NAO robot is available at: http://perso.ensta-paristech.fr/~tapus/eng/media.html. The video is composed of four scenes recognizing three emotions belonging to the existing clusters in the database (Figure 2-6), in addition to one new emotion not in-

New Data	Uncertain New Data (Scenario 2)	New Data Belonging to Old Data Clusters (Scenario1)						
		Anger	Disgust	Happiness	Sadness	Surprise	Fear	Neutral
Anxiety	81.6%	0	0	0	2.5%	0	15.9%	0
Shame	73.3%	0	13.3%	0	0	6.7%	0	6.7%
Desperation	68.75%	0	12.5%	0	6.25%	0	12.5%	0
Pride	73.3%	0	0	0	6.7%	6.7%	0	13.3%
Contempt	62.5%	6.25%	0	0	6.25%	0	18.75%	6.25%
Interest	75%	0	0	0	6.25%	6.25%	6.25%	6.25%
Elation	68.75%	6.25%	0	12.5%	0	0	0	12.5%
Boredom	69.8%	0	0	0	5.2%	0	23.9%	1.1%

Table 2.3 – Confusion matrix for the classification of the new data elements as being uncertain-emotion elements or as being a part of the existing clusters

cluded in the database. These emotions are: surprise, anger, boredom, and shame. The voice signal was acquired through a wireless ear microphone (hidden from the angle of the video camera).

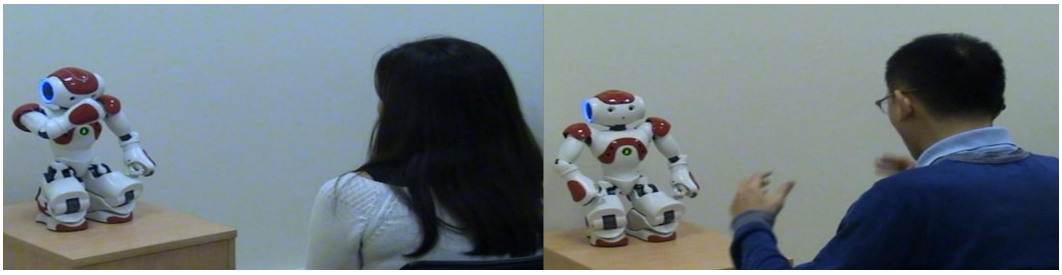


Figure 2-6 – Two participants are interacting with the robot. Each one expressed an emotion and the robot tried to recognize it. This recognition represents an action that the robot could generate corresponding to the expressed emotion.

The “surprise" and “anger" emotions were recognized successfully due to their distinguished vocal patterns. Meanwhile, the “boredom" emotion was confused with the “sadness" emotion due to the similarity between their vocal patterns, which made their recognition scores close to each other. Last but not least, the “shame" emotion was recognized correctly as a new emotion (i.e., not included in the database), after some confusion with one of the previously learnt emotions “anxiety". In the beginning, the expressed “shame" emotion to the robot was not attributed with a confident score to any of the existing classes. However, the “anxiety" class was the nearest winner class, but the attained score was less than 80%. Therefore, Scenario 2 (Section 2.6) was implemented. The expressed emotion was attributed to all the existing clusters, to which some data elements from the “shame" emotion class had been added, as if they represent the previously attributed uncertain data

to all the existing clusters. The objective was to find out to what extent the proposed algorithm would be able to detect a new emotion and - consequently - construct a corresponding new cluster.

An interesting future extension of this work that requires an elaborate study, is studying the performance of the proposed fuzzy-based emotion detection system within other dimensional continuums of emotion, in which an emotional state is characterized by a small number of latent dimensions, which were discussed in details in several studies [Russell, 1980; Scherer, 2000; Scherer et al., 2001; Fontaine et al., 2007; Grandjean et al., 2008]. These different approaches to addressing the multidimensionality of emotion would help in considering the nature of emotion during human-robot interaction so as to make the robot able to generate an appropriate behavior to the context of interaction.

2.8 Conclusions

This chapter illustrates an online detection approach of human's emotion using fuzzy logic. Our approach is based on the subtractive clustering algorithm that calculates the cluster centers of a data space. These centers represent the rules of the TS fuzzy models that characterize emotion clusters separately. Decisive criteria based on a recursive potential calculation for the new data decide whether the new elements constitute a new cluster, or they belong to one of the existing clusters. In case a new cluster is set up, a corresponding TS fuzzy model will be created. Meanwhile, in case the new data is attributed to one of the existing clusters, it may update the TS model of the winner cluster whether by creating a new rule, or by replacing one of the existing rules according to its descriptive power.

The obtained recognition scores of emotions in the offline system, prove the pertinence of the chosen prosody features in our analysis, upon which the online fuzzy-based emotion detection system is created. This fuzzy model proved its reasonable precision in detecting emotion online, determining if a new cluster needs to be created, and calculating if the new data elements are sufficiently descriptive to update an existing TS model.

When an uncertain-emotion data element is detected, or a new cluster is created, the robot

performs a neutral action at the beginning in order to avoid any inconsistency in the context of interaction. Progressively, the robot's experience will increase, which helps it create autonomously a relevant multimodal behavior by its own. This last point is presented in Appendix (C) as a future research direction for this work, which has been published in Aly and Tapus [2012e, 2015b].

Chapter 3

Generating an Adapted Verbal and Nonverbal Combined Robot's Behavior to Human's Personality

In Human-Robot Interaction (HRI) scenarios, an intelligent robot should be able to synthesize an appropriate behavior adapted to human's profile. Recent researches discussed the effect of personality traits on the verbal and nonverbal behaviors, which play a major role in transferring and understanding messages within the interaction. The dynamic characteristics of the generated gestures and postures during the nonverbal communication can differ according to personality traits, which similarly can influence the verbal content of human's speech. The presented research in this chapter tries to map human's verbal behavior to a corresponding robot's verbal and nonverbal combined behavior based on the extraversion-introversion personality dimension, measured through a psycholinguistic analysis of human's speech. We explored the human-robot personality matching aspect, in addition to the differences between the adapted combined robot's behavior expressed through speech and gestures, and the adapted speech-only robot's behavior within an interaction. Experiments with NAO robot are reported.

3.1 Introduction

Creating a socially intelligent robot able to interact with humans in a natural manner and to synthesize appropriately comprehensible multimodal behaviors in a wide context of interaction, is a hard task. This requires a high level of multimodal perception, so that the robot should understand the internal states, intentions, and personality dimensions of the interacting human in order to be able to generate an appropriate verbal and nonverbal combined behavior to the context of interaction.

The literature reveals hard efforts aiming to support the natural human-robot conversational interaction. Grosz [1983] tried to create a limited verbal natural language interface in order to access information in a database. An interesting theoretical study on the Natural Language (NL) was discussed in Finin et al. [1986], in which they tried to study the effect of using natural language interaction of rich functionality (e.g., paraphrasing, correcting misconceptions, etc.) on the effective use of expert systems. Another interesting theoretical study was discussed in Wahlster and Kobsa [1989] and Zukerman and Litman [2001], where they focused on the field of user modeling (i.e., understanding the user's beliefs, goals, and plans) in artificial intelligence dialog systems, and illustrated the importance of such modeling on interaction. Later on, some researches tried to depict how believable will be the dialogue systems that are adapted to the user's model (including the ability to explicitly and dynamically change the aspects of the relationship with the interacting human through the use of social talks in the same way as humans behave) [Andre et al., 2000; Cassell and Bickmore, 2003; Forbes-Riley and Litman, 2007; Forbes-Riley et al., 2008].

Furthermore, some efforts were driven towards generating synchronized verbal and nonverbal behaviors, as discussed in Ng-Thow-Hing et al. [2010]. The authors presented a system able to synchronize expressive body gestures with speech. This model was implemented on Honda humanoid robot (ASIMO), and was able to synthesize gestures of different types, such as: iconic, metaphoric, deictic, and beat gestures [McNeill, 1992, 2000]. Moreover, Le and Pelachaud [2012] discussed an interesting system for synthesizing co-speech and gestures for NAO robot. They used the SAIBA framework [Kopp et al., 2006] in order to

generate a multimodal behavior designated to virtual agents, then they interfaced it with NAO robot in order to generate and model a synchronized verbal and nonverbal combined robot's behavior. Similarly, virtual agents had received much attention concerning generating expressive behaviors. Kopp et al. [2008] tried to simulate the natural speech-gestures production model that humans have on the 3D agent MAX. They proposed an architecture for generating synchronized speech and gestures in a free and spontaneous manner. For example, it is sufficient to support the system with some a priori information about a certain object to describe, and the system will be able to generate itself an expressive verbal and nonverbal combined behavior as humans do. Another interesting approach was discussed in Hartmann et al. [2002], Bevacqua et al. [2004], Mancini and Pelachaud [2008], and Niewiadomski et al. [2009]. The authors developed the virtual conversational agent GRETA, which uses verbal and nonverbal behaviors to express intentions and emotional states. It can be used as a dialog companion, a virtual tutor, a game-actor, or even a storyteller. Cassell et al. [2000] introduced the conversational agent REA, which presents a real estate sales person through a multimodal expressive behavior. Courgeon et al. [2008] introduced the multimodal affective and reactive character MARC, which can generate an expressive behavior in real time corresponding to the user's action. Despite the rich literature of generating expressive behaviors with robots and 3D agents, and to the best of our knowledge, no research work discussed the importance of generating a combined verbal and nonverbal robot's behavior based on the interacting human's personality traits.

Personality is an important factor in human social interaction. In the literature, there are different models of personality, such as: Big5 (*Openness, Conscientiousness, Extraversion-Introversion, Agreeableness, and Neuroticism*) [Goldberg, 1990, 1999], Eysenck Model of Personality (PEN) (*P: Psychoticism, E: Extraversion, and N: Neuroticism*) [Eysenck, 1953, 1991], and Meyers-Briggs (*Extraversion-Introversion, Sensation-Intuition, Thinking-Feeling, and Judging-Perceiving*) [Myers-Briggs and Myers, 1980; Murray, 1990]. In this research, the Personality Recognizer toolkit (Section 3.3.1) integrated to our system is based on the Big5 personality model, as it is the most descriptive model of human's personality. Morris [1979], Dicaprio [1983], Woods et al. [2005], and Tapus and Matarić [2008] defined personality as: "*the pattern of collective character, behavioral, temper-*

amental, emotional and mental traits of an individual that has consistency over time and situations". Consequently, it is obvious that the long term effect of personality on the generated behavior, makes it more reliable for characterizing the generated verbal and nonverbal behaviors, to the contrary of other short-term characteristics, like the prosodic features.

Based on these findings, we assume that personality is an important factor within a human-robot interaction context. In this research, we try to develop a customized robot's verbal and nonverbal combined behavior based on the extraversion-introversion personality trait of the interacting human. We focus on validating that the participants prefer more interacting with the robot when it has a similar personality to theirs, and that the adapted multimodal robot's combined behavior (i.e., robot-user personalities match in terms of the type and the level of the extraversion-introversion dimension, and that both speech and gestures are expressed synchronously) is more engaging than the adapted speech-only robot's behavior (not accompanied with gestures). The context of interaction in this research is restaurant information request, in which the robot gives the required information about restaurants to the interacting human in real-time, expressed through a combined verbal and nonverbal behavior [Aly and Tapus, 2012a, 2013a].

The rest of the chapter is structured as following: Section (3.2) discusses the importance of personality traits in human-robot interaction, Section (3.3) presents a general overview of the system architecture, Section (3.4) describes the nonverbal behavior's knowledge base extension, Section (3.5) illustrates how we realized the synchronized verbal and nonverbal behaviors on the robot, Section (3.6) illustrates the design, the hypotheses, and the scenario of interaction, Section (3.7) provides a description of the experimental results, Section (3.8) discusses the outcome of the study, and finally, Section (3.9) concludes the chapter.

3.2 Why should personality traits be considered in human-robot interaction?

In human-robot interaction, a straightforward relationship has been found between personality and behavior [Nass and Lee, 2001; Eriksson et al., 2005; Woods et al., 2007]. In the

context of human modeling and adapting the dialog of a machine (i.e., a humanoid robot or a computer) to the personality of the interacting human, Reeves and Nass [1996], Nass and Lee [2001], and Tapus and Matarić [2008] proved empirically that the human interacting with a dialog machine will spend more time on the assigned task if the system's behavior matches with his/her personality, which validates the similarity attraction principle (i.e., individuals are more attracted by others who have similar personality traits) in human-robot interaction situations [Byrne and Griffit, 1969]. Another interesting topic was discussed in Park et al. [2012], in which they examined the influence of the KMC-EXPR robot's personality (reflected only through facial expressions using the eyes and the mouth, with big movements for extraverts and small movements for introverts) on its anthropomorphism, friendliness, and social presence. The results showed that the participants assigned the extraverted robot a higher degree of anthropomorphism than the introverted robot. On the other hand, for the friendliness and social presence, the results depicted that the extraverted participants considered the extraverted robot more friendly and socially present than the introverted robot, while the introverted participants preferred more the introverted robot. These findings validate the similarity attraction principle [Byrne and Griffit, 1969].

Another interesting concept is the complementarity attraction (i.e., individuals are more attracted by others whose personalities are complementary to their own personalities) [Sullivan, 1953; Leary, 1957; Isbister and Nass, 2000]. The effect of the AIBO robot's personality on the interacting participants through relatively long-duration experiments, has been studied in Lee et al. [2006]. The authors found that the participants preferred interacting more with the robot when it had a complementary personality than when it had a similar personality, to their own personalities. Generally, the confusion between the similarity and complementarity attraction principles could be related to the context of interaction, so that any of them could be validated during a human-robot interaction experiment, similarly to the human-human social attraction that involves either the similarity or the complementarity attraction during interaction [Dijkstra and Barelds, 2008]. For example, the similarity attraction looks more appropriate for the experimental design that considers the effect of the initial interaction between human and robot on the developing relationship (which could be figured in most friendships between humans, where they get attracted to each other based

on the matching between their personalities, and on the equality of dominance between each other). Meanwhile, the complementarity attraction contends more for long-term relationships (e.g., marriage and some kinds of friendship of different roles, where one person is more dominant than the other) [Vinacke et al., 1988].

In this research, we are interested in making the interacting human more attracted to the robot during the conducted experiments, so that the robot takes a similar personality to the interacting human's personality (i.e., similarity attraction principle is being examined). Furthermore, due to the relatively short-duration of the conducted experiments, the validation of the complementarity attraction principle (using the current experimental design) would be hard to be accomplished.

A strong psychological evidence that firmly supports our focus on the similarity attraction principle, is the *chameleon effect*. This effect refers to the "*nonconscious mimicry of the postures, mannerisms, facial expressions, and verbal and nonverbal behaviors of one's interaction partners, such that one's behavior passively and unintentionally changes to match that of others in one's current social environment*", which happens frequently and naturally between people [Chartrand and Bargh, 1999]. This definition matches the findings of Bargh et al. [1996], which suggested that the perception of one's behavior enhances the chances of engaging in that behavior by his/her counterpart. Giles and Powesland [1978] discussed mimicry in speech and found that people tend to mimic the accents of their interaction partners. Other speech characteristics like speech rate and rhythm are also mimicked during interaction [Webb, 1972; Cappella and Planalp, 1981].

Similarly, Lafrance [1982] and Bernieri [1988] found that gestures, postures, and mannerisms are mimicked during interaction. This verbal and nonverbal behavior mimicry reported a higher positive effect on interaction than the cases when mimicry was absent [Chartrand and Bargh, 1999]. Maurer and Tindall [1983] found that the mimicry of a client's arm and leg positions by a counselor, increased the client's perception of the empathy level of the counselor. Van-Baaren et al. [2003] found that when a waitress mimicked verbally her customers, she received a larger amount of tips. Bailenson and Yee [2005] found that mimicking the participant's head movements by a virtual agent was per-

ceived more convincing and was attributed a higher trait ratings than the non-mimicking interaction cases. Moreover, several studies investigated the relationship between behavior mimicry and attraction. Gump and Kulik [1997] discussed that behavior mimicry enhances the coherence within interaction by making the interacting partners look similar to each other. Gueguen [2007] studied the effect of the verbal and nonverbal behavior mimicry on a courtship relationship. He found that the male participants preferred the female participants who mimicked them. Luo et al. [2013] found that people preferred similar gestures to their own during human-agent interaction, which matches the outcome of the previous studies. Additionally and most importantly, this study suggested a *preliminary* relationship between personality and the perception of an exercised behavior. This last primary result, in addition to all the previous discussion open the door in front of a more elaborate study that investigates the link between personality and behavior, which constituted a strong inspiration for our current work.

Barrick and Mount [1991] investigated the general relationship between personality and professions. They found that some professions, such as: teacher, accountant, and doctor, tend to be more introverted, while other professions, such as: salesperson and manager, tend to be more extraverted. A similar tendency was discussed in Windhouwer [2012], which tried to investigate how could NAO robot be perceived intelligent in terms of its profession and personality. They found that when the robot played the role of an introverted manager, it appeared more intelligent than the extraverted manager. Similarly, when the robot played the role of an extraverted teacher, it appeared more intelligent than the introverted teacher. These last findings oppose - to some extent - the findings of Barrick and Mount [1991], which could be due to some differences in the context of interaction. For example, when the robot was playing the role of an introverted manager during a meeting, it probably seemed deeply thinking about work problems trying to reach optimal solutions. This could have given the introverted robot a more intelligent look than the extraverted robot that was not looking thinking enough, and was moving fast with high energy. Therefore, the findings of Barrick and Mount [1991] could be considered as general findings that could differ experimentally according to the context of interaction, which makes the matching between the robot's personality and profession (task), a difficult point to esti-

mate in advance before experiments. However, it is worthy with study, as it can influence positively the way people perceive the robot.

Moreover, Leuwerink [2012] discussed how people would perceive the robot intelligent in terms of its personality within dyadic and group interactions. They found that the introverted robot was perceived more intelligent in a group interaction. Meanwhile, the extraverted robot was perceived more intelligent in a dyadic interaction. These findings match the findings of Barrick and Mount [1991] in a general manner for certain professions, such as: teacher for the introverted robot in a group interaction, and salesperson for the extraverted robot in a dyadic interaction. However as mentioned earlier, it all depends on the context of interaction, because an extraverted robot-teacher could be more suitable for a group interaction, considering that it will appear more active and funny. Therefore, it is difficult to draw a common and general definition for the relationship between personality, profession, and group/dyadic interaction due to the differences that may appear in each experimental study.

On the other hand, other studies found a relationship between human's personality and proxemics (i.e., the study of the interpersonal distance's influence on interaction) [Hall, 1966; Tapus et al., 2008], which influences the robot's navigation planners in human-robot interaction situations (e.g., extraverted people are more tolerant of their personal space invasion by a robot than introverted people) [Williams, 1971]. Nakajima et al. [2003, 2004] discussed the influence of emotions and personality on the social behaviors of human-robot collaborative learning systems. They found that the users had more positive impression about the usefulness of the learning experience when the cooperative agent displayed some social responses with personality and emotions. Generally, all the previous discussion reveals the feasibility of considering personality traits in human-robot interaction scenarios, which can attract humans to interact more efficiently with robots.

Previous researches discussed the importance of the extraversion-introversion dimension in characterizing human's behavior. J-Campbell et al. [2003] and Selfhout et al. [2010] discussed the important effect of both the agreeableness and the extraversion-introversion dimensions on developing human peer relationships. Lippa and Dietz [2000] indicated that

the extraversion-introversion dimension is the most influential and accurate trait among the Big5 personality dimensions. Besides, Moon and Nass [1996], Isbister and Nass [2000], and Nass and Lee [2001] discussed the importance of the extraversion-introversion dimension in Human-Computer Interaction (HCI). On the other hand, several researches considered the verbal and nonverbal cues as the most relevant cues for personality traits analysis [Riggio and Friedman, 1986; Pittam, 1994; Hassin and Trope, 2000; Nass and Lee, 2001]. Consequently, this work tries to demonstrate the influence and the importance of personality in human-robot interaction contexts. It links between the extraversion-introversion dimension and the verbal and nonverbal behavioral cues for the purpose of generating an adapted robot's behavior to human's personality so as to reinforce the level of interaction between human and robot.

3.3 System architecture

Our system is a coordination between different sub-systems: (1) Dragon naturally speaking toolkit, which translates the spoken language of the interacting human into a text, (2) Personality Recognizer toolkit, which estimates the interacting human's personality traits through a psycholinguistic analysis of the input text [Mairesse et al., 2007], (3) PERSONAGE natural language generator, which adapts the generated text to the interacting human's personality dimensions [Mairesse and Walker, 2011], (4) BEAT toolkit, which translates the generated text into gestures (not including the general metaphoric gestures) [Cassell et al., 2001], (5) Metaphoric general gesture generator (Section 3.5) (which will be explained in details in Chapter 4) [Aly and Tapus, 2013b], and (6) NAO robot as the test-bed platform. An overview of the system architecture is illustrated in Figure (3-1).

3.3.1 Personality Recognizer

Personality markers in language had received a lot of interest from psycholinguistic studies. Scherer [1979], Furnham [1990], and Dewaele and Furnham [1999] described how could the extraversion-introversion personality trait influence linguistically speech production. They stated that extraverts are more loud-voiced, and talk more iteratively with less falter-

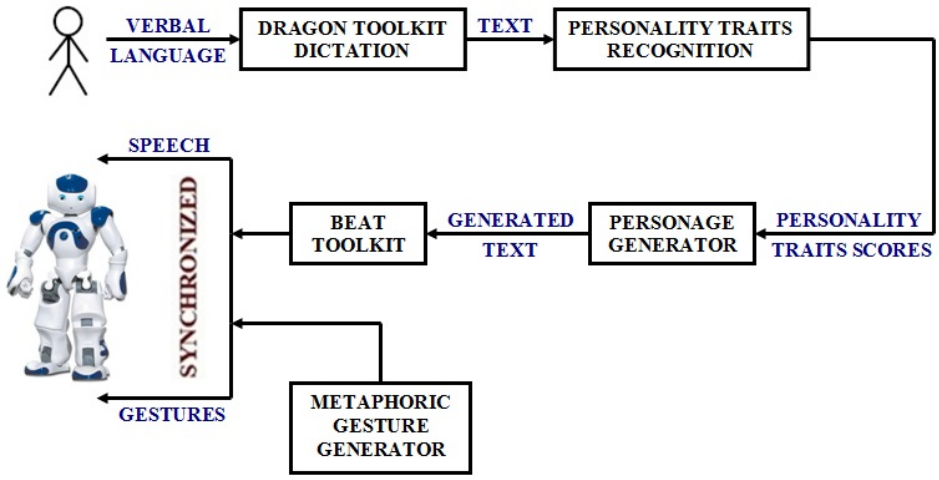


Figure 3-1 – Overview of the adapted verbal and nonverbal combined behavior generating system architecture

ing and pauses than introverts. Moreover, extraverts have a higher verbal output, informal language, and speech rates, while introverts use a richer vocabulary. On the other hand, extraverts express more encouragement and agreement, and use positive feeling words more than introverts [Pennebaker and King, 1999].

A general approach for characterizing the majority of personality traits was discussed in Pennebaker and King [1999], in which they used the Linguistic Inquiry and the Word Count toolkit (LIWC) in order to define the word categories of 2479 essays (containing 1.9 million words) written by different persons covering the five personality traits described in the Big5 Framework [Goldberg, 1990, 1999]. This dictionary enabled them to state general relationships and characteristics for the five personality traits. Conscientious people -for example- avoid negative feeling words, negations, and words expressing discrepancies. Similarly, Mehl et al. [2006] created a spoken data corpus (containing 97468 words and 15269 utterances), in addition to their transcripts, covering different personality traits. This corpus was sub-divided into several word categories using the LIWC tool.

The findings of the previous data corpora were the basic body of the research conducted by Mairesse et al. [2007]. They created a huge database including the LIWC psycholinguistic features, such as: anger words (e.g., hate), metaphysical issues (e.g., god), and family members (e.g., mom, and brother), in addition to other psycholinguistic features included in

the MRC database [Coltheart, 1981], such as: frequency of use (e.g., low: nudity, duly, and high: the, he) and concreteness (e.g., low: patience, and high: ship), besides the utterance type features, such as: command (e.g., must, and have to), prompt (e.g., yeah, and ok), and question-assertion (which is any utterance out of the previous categories). The relationship between the utterance type features and personality traits was discussed in Vogel and Vogel [1986] and Gill and Oberlander [2002], in which for example, extraverts are more assertive when writing emails. Afterwards, the system was trained on the previously stated data corpora using the Support Vector Machines (SVM) algorithm, and was cross validated so as to approve its performance.

3.3.2 PERSONAGE Generator

PERSONAGE is a natural language generator that can express several personality dimensions through language. The architecture of PERSONAGE generator is illustrated in Figure (3-2), which is based on the traditional pipelined natural language generation (NLG) architecture [Reiter and Dale, 2000]. The input consists of personality traits' scores, besides the selected restaurant(s) in New York City. The database of PERSONAGE generator contains scalar values representing the ratings of 6 attributes (used for recommendation and/or comparison according to the experimental context): cuisine, food quality, service, atmosphere, price, and location, of more than 700 restaurant collected from real surveys investigating the opinion of people visited these restaurants. The content of the generated language could be more controlled through some parameters, like the verbosity parameter, which could be set to 1 in order to maximize the wordy content of the generated utterance.

The content planner plays the role of choosing and structuring (in a tree format) the necessary information to be processed by the sentence planner, in terms of the values of some parameters, such as: verbosity, polarity, and repetition (i.e., the content planer decides what to say). Meanwhile, the sentence planner deals with phrasing the information structured by the content planner. It searches in the dictionary, the group of primary linguistic structures attributed to each proposition in the content plan (e.g., if the content planner structured a recommendation, the sentence planner would precise the syntactic parts of the recommen-

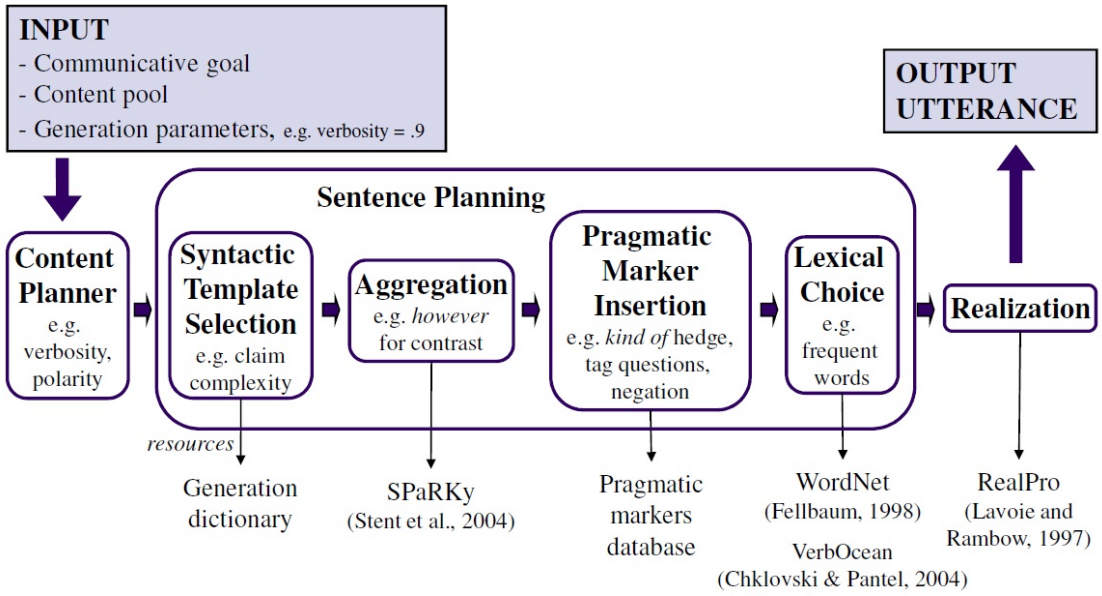


Figure 3-2 – Architecture of PERSONAGE generator [Mairesse and Walker, 2011]

dation, such as: verb, noun, etc.). Afterwards, it aggregates the obtained syntactic templates in order to generate a complete syntactic structure for the utterance [Stent et al., 2004].

On the other hand, the pragmatic marker insertion process in the sentence planner modifies the aggregated syntactic structure in order to generate several pragmatic effects, like: the hedge *you know*, the question tags, etc. The lexical choice process chooses the most appropriate lexeme from many different lexemes expressed by PERSONAGE generator, for each word in terms of the frequency of use, the length, and the lexeme's strength [Fellbaum, 1998; Chklovski and Pantel, 2004]. Last but not least, the realization process follows the sentence planner and transforms the resulting syntactic structure to a string using appropriate rules (i.e., the word insertion and morphological inflection rules) [Lavoie and Rambow, 1997].

3.3.3 BEAT Toolkit

BEAT is the Behavior Expression Animation Toolkit that takes as an input a text and generates a corresponding synchronized set of gestures. It processes the contextual and linguistic information of the text so as to control body and face gestures, besides voice intonation. This mapping (from text to gesture) is implemented through a set of rules derived from

intensive research on the nonverbal conversational behavior [Cassell et al., 2001]. BEAT pipeline is composed of different XML-based modules, as illustrated in Figure (3-3). The language tagging module receives an XML tagged text generated from PERSONAGE generator, and converts it into a parse tree with different discourse annotations (e.g., theme and rheme). The behavior generation module uses the output tags of the language module and suggests all possible gestures, then the behavior filtering module selects the most appropriate set of gestures using the gesture conflict and priority filters. The user-definable data structures, like: the generator, and filter sets (indicated in dotted lines), provide the generation and filtering rules and conditions for the behavior generation and selection processes. Meanwhile, the knowledge base adds some important contextual information and definitions for generating relevant and precise nonverbal behaviors, such as:

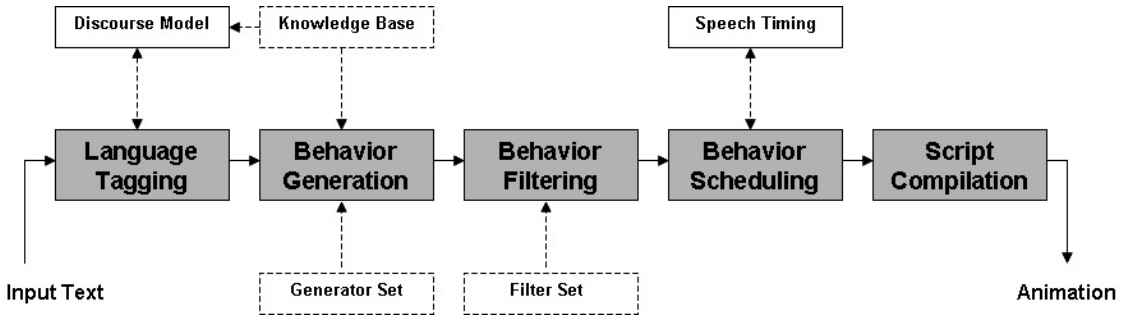


Figure 3-3 – Architecture of BEAT toolkit [Cassell et al., 2001]

– **T**ype: which attributes features with their values to different object types (e.g., the object “Home", which belongs to the class “Place" with type features attributes as “House, Apartment").

– **I**nstance: which describes specific cases of recognizable objects (e.g., the “Spiral" shape could be considered as a shape instance of the object “Stairs").

– **S**cene: which groups all instances of the same environment into scenes.

– **G**esture: which specifies different kinds of gestures and their proposed trajectories and hand shapes.

The behavior scheduling module converts the input XML tree into a set of synchronized speech and gestures. It includes a TTS (text-to-speech) engine that calculates the duration

of words and phonemes, which helps in constructing an animation schedule for the aligned gestures with words. The script compilation module compiles the animation script into some executive instructions that can be used in animating a 3D agent or a humanoid robot.

3.4 Extension of the nonverbal behavior knowledge base of BEAT toolkit

The purpose of the performed extension on BEAT toolkit was to add necessary information about the generated text by PERSONAGE generator comparing (and/or recommending) between different restaurants in New York City. The object-type “Restaurant" is defined as an object in the class “Place" with some information about the restaurant’s location, price category, size, and cuisine, which has been used in the scenarios of interaction. Some instances were also added to the knowledge base describing some related places to the object “Restaurant", such as: “Basement" and “Dining Room" in terms of their size, lightening, and painting. The new added scenes to the knowledge base define the restaurants’ names, including the previously defined instances. The precised gestures’ characteristics in the knowledge base concern different types of iconic gestures, including hand shapes and arm trajectories (unlike other gesture categories that do not require specific hand/arm shapes, as indicated in Section 3.5). Some new linguistic keywords were aligned to specific iconic gestures with the corresponding hand/arm geometrical shapes’ characteristics, like the adjective “big", which was aligned to the hand shape “hands-in-front" and the arm trajectory “big-span" in order to refer to a big span separating between the two hands, which semantically matches the adjective “big".

3.5 Modeling the synchronized verbal and nonverbal behaviors on the robot

BEAT toolkit was built as a customizable gesture generator, so that more gesture categories could be added to the generation system of the toolkit, or even some extension could be im-

posed on its nonverbal behavior knowledge base in order to increase the expressivity scope of some built-in gestures (e.g., iconic gestures), as indicated in Section (3.4). Generally, we found that the built-in gesture categories are mostly sufficient for the relatively short verbal context generated by PERSONAGE generator (except for the general metaphoric gestures, which are not included in BEAT toolkit. Therefore, we have integrated them externally to the system, as illustrated in Figure 3-5). In this research, we are interested only in four categories of gestures: iconic, posture-shift, metaphoric, and gaze gestures.

The animation script (generated by BEAT toolkit) described in Figure (3-4), indicates the proposed synchrony between the verbal content and the corresponding allocated gestures of the following sentence: *The first restaurant was calm and not far from downtown, but expensive. The second restaurant had not only a big dining room, but also a better quality and it was cheaper*. The system divides the sentence into chunks, where each chunk contains a group of words with specific allocated gestures. The symbol WI indicates the index of words (31 words in total), while the symbol SRT defines the estimated duration of each group of words with the allocated gestures. The animation script reveals also that the adjective word “big" was attributed to an iconic gesture, where the two hands are used to depict the gesture “gesture-both-hands" (i.e., performing a gesture using both hands) with the shape “hands-in-front", which proves the importance of customizing the knowledge base in order to generate the most appropriate nonverbal behavior.

Metaphoric gestures (which are not present in the animation script in Figure 3-4) are used frequently in order to represent the narrated speech but not in a physical manner, like iconic gestures. They could take the form of a general hand/arm/head shaking or even a specific shape, like when we want to express a time sequence, we use the word "after" associated with a specific hand/arm motion symbolizing this idea. Therefore, this word and other similar new words, were added and allocated in the knowledge base to the corresponding specific hand/arm motion trajectory, similarly to iconic gestures. On the other hand, the generation of general metaphoric gestures does not follow a specific linguistic rule, which makes it a virtual generation of gestures. Our approach associates the generation of general metaphoric gestures to some prosodic rules so as to integrate the paraverbal modality into

```
<?xml version="1.0" encoding="UTF-8"?>
- <AnimationScript HEARER="USER" SPEAKER="AGENT">
    <START SRT="0.0" WI="0" SPEECH=" The first restaurant was calm and not far from downtown, but expensive.
      The second restaurant had not only a big dining room, but also a better quality and it was cheaper.
    " ACTION="SPEAK" AID="A527"/>
    <START SRT="0.0" WI="0" ACTION="GAZE" AID="A535" PRIORITY="1" DIRECTION="AWAY_FROM_HEARER"/>
    <START SRT="0.0" WI="0" ACTION="POSTURESHIFT" AID="A539" ENERGY="LOW" BODYPART="BOTH"/>
    <STOP SRT="1.01" WI="1" ACTION="POSTURESHIFT" AID="A539" ENERGY="LOW" BODYPART="BOTH"/>
    <STOP SRT="1.52" WI="3" ACTION="GAZE" AID="A535" PRIORITY="1" DIRECTION="AWAY_FROM_HEARER"/>
    <START SRT="1.52" WI="3" ACTION="GAZE" AID="A552" PRIORITY="5" DIRECTION="TOWARDS_HEARER" FOCUS="ANY"/>
    <STOP SRT="7.47" WI="19" ACTION="GAZE" AID="A552" PRIORITY="5" DIRECTION="TOWARDS_HEARER" FOCUS="ANY"/>
    <START SRT="7.47" WI="19" ACTION="GAZE" AID="A552" PRIORITY="5" DIRECTION="TOWARDS_HEARER" FOCUS="ANY"/>
    <STOP SRT="7.54" WI="20" ACTION="GESTURE_BOTH_HANDS" AID="A563" PRIORITY="20" TRAJECTORY="BIG_SPAN"
      SHAPE="HANDS_IN_FRONT"/>
    <START SRT="7.54" WI="20" ACTION="GAZE" AID="A552" PRIORITY="5" DIRECTION="TOWARDS_HEARER" FOCUS="ANY"/>
    <STOP SRT="11.54" WI="30" ACTION="GAZE" AID="A552" PRIORITY="5" DIRECTION="TOWARDS_HEARER" FOCUS="ANY"/>
</AnimationScript>
```

Figure 3-4 – Synchronization XML animation script for the generated verbal and nonverbal combined behavior

the generation of a nonverbal behavior, as will be explained in details later on (Chapter 4).

The mapping of gaze, posture-shift, iconic, and specific-shape-metaphoric gestures to the robot from the animation script (Figure 3-4) necessitates that the robot processes each line of the script indicating the duration of each chunk that contains a synchronized verbal content with an attributed nonverbal behavior. Kendon [1980] defined *gesture phrases* as the primary units of gestural movement that include consecutive movement phases, which are: *preparation*, *stroke*, and *retraction* beside some intermediate *holds*. The problem that may appear when modeling a combined verbal and nonverbal behavior on the robot (in case of the iconic and specific-shape-metaphoric gestures), is the required high temporal synchronization between the stroke (i.e., the expressive gestural phase) and the affiliate (i.e., the affiliated word or sub-phrase), in order to express an idea accurately. The time estimation indicated in the animation script reveals the calculated time for the stroke phase of gesture. Consequently, an additional time estimation for the preparation phase should be assumed, so that the hands/arms leave their initial position and get ready for the stroke phase synchronously with the affiliate. Therefore, the gesture's stroke phase is fixed to lead the affiliate's onset by an approximate duration of one syllable (i.e., 0.3s).

Figure (3-5) illustrates the robot's gestural behavior control architecture. The general metaphoric gesture generator receives as an input, the temporally aligned text with speech using a TTS (text-to-speech) engine, so that it synthesizes general metaphoric gestures corresponding to each word of the text based on the prosodic cues of the aligned speech segments to words [Tapus and Aly, 2011; Aly and Tapus, 2011a,b, 2012c,d,b, 2013b]. Conse-

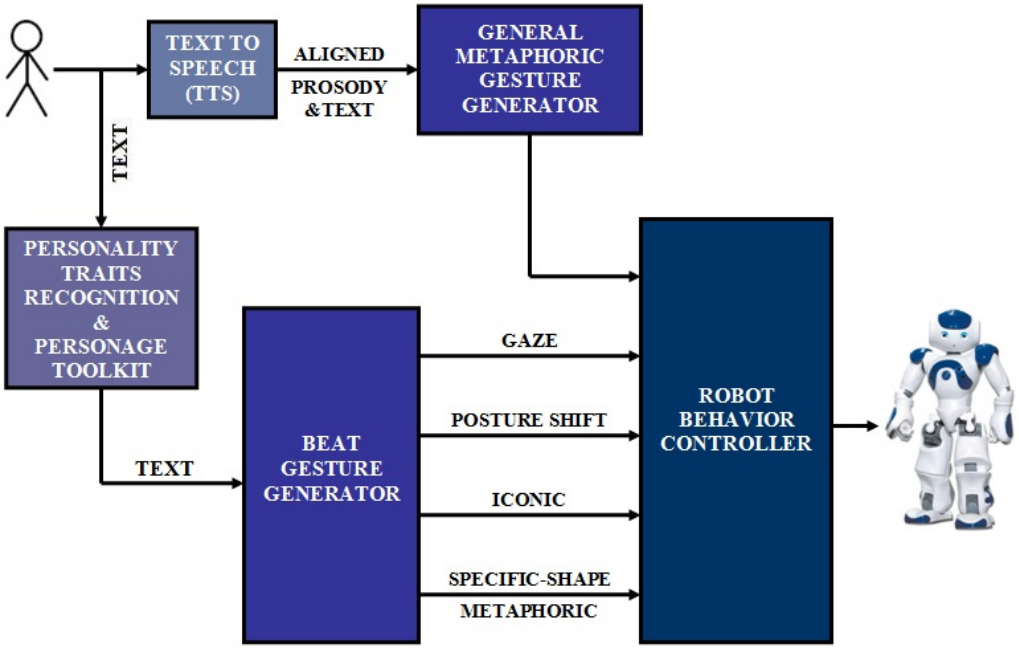


Figure 3-5 – Robot's gestural behavior control

quently, the final chunks in the behavior controller would contain both the temporal and the corresponding word-index information of five gesture types (i.e., general metaphoric gestures, gaze gestures, posture-shift gestures, iconic gestures, and specific-shape-metaphoric gestures).

General metaphoric gestures are synthesized using the Coupled Hidden Markov Models (CHMM), which could be considered as a multi-stream collection of parallel HMM characterizing the segmented data of both prosody and gestures. The generated gestures are characterized by the most likely path of observations through the gesture chain of the CHMM (which is modeled in terms of the linear velocity and acceleration observations of body segments, in addition to the position and displacement observations of body articulations), given an observed audio sequence [Aly and Tapus, 2013b] (this part will be explained in details in Chapter 4). Having known the position coordinates and orientation of the synthesized gestures, the inverse kinematics is applied in order to calculate the corresponding rotation values of body articulations so as to get modeled on the robot (with the help of the synthesized velocity, acceleration, and displacement motion curves).

Using the CHMM in generating metaphoric gestures allows synthesizing gestures of vary-

ing amplitude and duration. Besides, the random variations of the synthesized gestures' motion patterns make them look as natural as human gestures. This methodology clarifies the quantitative difference between the generated amount of gestures in case of the introverted and extraverted conditions. Therefore, for an introverted speaker who does not speak a lot, he/she will have a corresponding limited pitch-intensity contours, which will lead to a corresponding limited set of generated gestures, contrarily to the extraverted individuals.

On the other hand, in order to reasonably reflect a specific introverted or extraverted personality on the robot, the generated motion curves' values of the synthesized gestures should be controlled in both personality conditions. Consequently, we attributed experimentally 10% of the amplitude of the generated motion curves' values to the maximum introversion level, while we kept 100% of the amplitude for the maximum extraversion level (based on the fact that the training database for the CHMM was based on highly extraverted actors) [Aly and Tapus, 2013b]. The corresponding motion curves' values to the range of personality scores between the maximum introversion and extraversion levels (i.e., between 10% and 100%) could be easily derived as a function of the motion curves' values calculated at the maximum introversion and extraversion levels.

Unlike the automatic modeling of the synthesized general metaphoric gestures on the robot directly, the modeling of the other four types of gestures generated by BEAT toolkit was controlled inside the robot's behavior controller. During the gaze gesture (whether it is oriented *towards the hearer* or *away from the hearer*), the whole neck turns so as to get oriented away/towards the interacting human (Figure 3-4). The neck movement was previously programmed (same for the posture-shift gesture in the directions: *lean forward* and *lean backward*). A similar tendency was applied for the generated iconic and specific-shape-metaphoric gestures, in which corresponding body movements to certain words in the knowledge base were also previously programmed. The motion control parameters of the generated gestures by BEAT toolkit have initially been set experimentally through the normal range of personality scores, from 10% (maximum introversion) to 100% (maximum extraversion) with a step of 10%, so that the robot implements the generated gestures in a corresponding approximate manner to the desired personality type and level to show. How-

ever, the encountered difficulty was to keep the temporal alignment between the generated gestures and text indicated in the animation script in Figure (3-4). Therefore, the robot's behavior controller should be updating the time-control parameter of the programmed gestures based on their estimated duration in the animation script so as to make the robot finish performing a specific gesture at the specified time instants in the script.

After designing the nonverbal behaviors corresponding to the five gesture types previously explained, the robot's behavior controller, finally, examines any existing conflict between the synthesized gestures. If there exists a conflict between an iconic or a specific-shape-metaphoric gesture (less frequent) and a general-hand/arm-metaphoric gesture (more frequent), so that both have to be implemented at the same time, the priority would be given automatically to the iconic or the specific-shape-metaphoric gesture. A similar tendency happens if a conflict occurs between a gaze gesture (in the direction *away from the hearer*) and a general-head-metaphoric gesture, in which the priority goes to the gaze gesture.

3.6 Experimental setup

In this section, we introduce the experimental hypotheses, the design, and the scenario of interaction between the participant and the humanoid NAO robot developed by Aldebaran Robotics (Section 1.2.1).

3.6.1 Hypotheses

The presented research aim to test and validate the following hypotheses:

- **H**1: The robot's behavior that matches the user's personality expressed through combined speech and gestures will be preferred by the user.

- **H**2: The robot's personality expressed through adapted combined speech and gestures will be perceived more expressive by the user than the robot's personality expressed only through adapted speech.

3.6.2 Experimental Design

In order to test and validate the first hypothesis, the user was exposed to two robot's personalities:

- The robot uses introverted cues expressed through combined gestures and speech, in order to communicate with the user.
- The robot uses extraverted cues expressed through combined gestures and speech, in order to communicate with the user.

Similarly, in order to validate the second hypothesis, the user tested two different conditions:

- The robot communicates with the user through combined gestures and speech (robot-user personalities match in terms of the type and the level of personality). We call it: **adapted combined robot's behavior**.
- The robot communicates with the user only through speech (robot-user personalities match in terms of the type and the level of personality). We call it: **adapted speech-only robot's behavior**.

All the previous four conditions were randomly ordered during the experimental phases. For the second hypothesis, we excluded the condition of interaction through gestures only, as it does not fit in the normal context of the non-mute human-human interaction. Similarly, we excluded the condition of interaction through adapted speech and non-adapted gestures to human's personality, because of the following reasons: (1) the production of human gestures and speech follows the same process, so that they are naturally aligned, and (2) the characteristics of the naturally aligned speech and gestures of human are adapted to his/her personality, therefore the robot's generated speech and gestures should be both adapted to the interacting human's personality so as to make the interaction more engaging. Consequently, it is neither normal nor natural to consider that speech could be adapted to human's personality alone without gestures (similarly to the adapted gestures and non-adapted speech interaction condition, which has not been considered in our study).

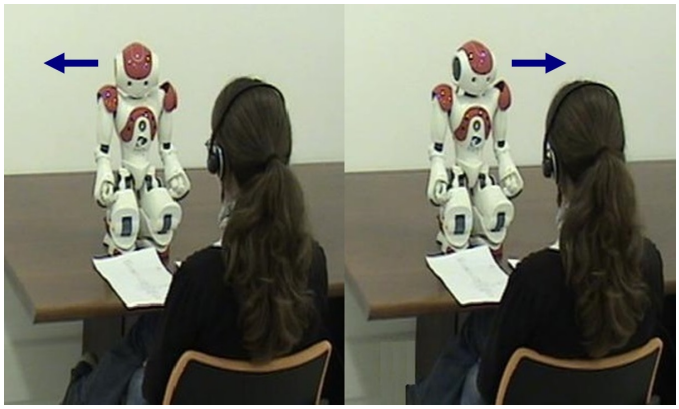


Figure 3-6 – Introverted robot condition (the robot's gaze was more down-directed with a low gestures rate. The arrows refer to the direction of head movement.)

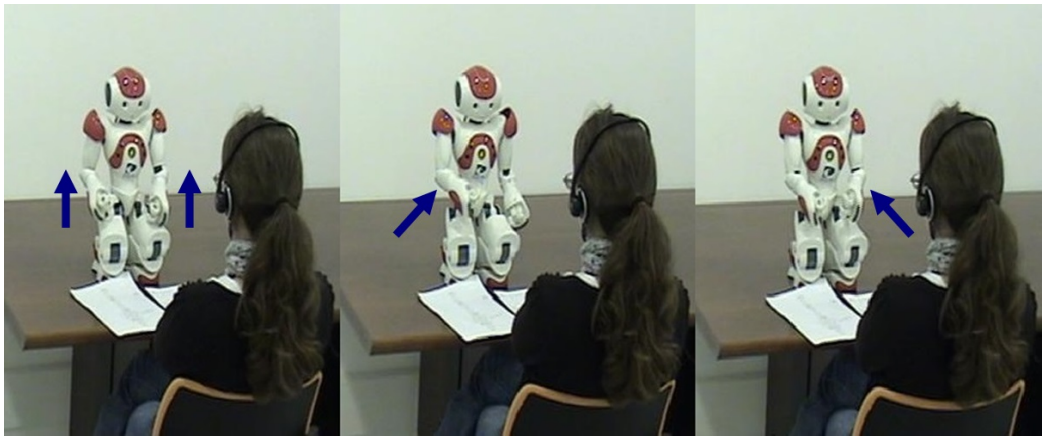


Figure 3-7 – Extraverted robot condition (the robot's head was looking more up, and the general metaphoric gestures rate of both head and arms was high. The arrows refer to the direction of arms' movement.)

Generally, the main objective of the second hypothesis is to evaluate the importance of using adapted combined speech and gestures together during communication (instead of using adapted speech only), in order to better express and reflect ideas. In our experiments, we focused only on the extraversion-introversion dimension that indicates the level of sociability of an individual. An extraverted individual tends to be sociable, friendly, fun loving, active, and talkative, while an introverted individual tends to be reserved, inhibited, and quiet (Figures 3-6 and 3-7).

The theme of the interaction in our experiments is restaurant information request. The robot has a list of restaurants in New York City, and its role is to give appropriate information about six elements: cuisine, food quality, service, location, atmosphere, and price, for the

selected restaurants in comparison. Our interaction scenario is described as following:

- The robot introduces itself as a guide to the participant, and asks him/her to say somethings he/she knows about New York City. This first step is necessary for the robot in order to be capable of automatically identifying the participant's personality based on the analyzed linguistic cues.
- The robot has a list of restaurants, and asks the participant to choose some restaurants so as to find out more details about them.
- The robot waits for the participant's input so as to produce appropriate combined speech and gestures based on the calculated personality traits.
- The participant asks for information about two restaurants of his/her choice.
- The robot gives the required information through a combined verbal and nonverbal behavior to the participant in real time.
- The participant can ask for more details about other restaurants (if he/she wants), and the robot gives back the required information.
- The interaction ends when the participant does not want to know more information about other restaurants, so that he/she has got the required information that was searching for.

The following examples indicate the differences between the generated verbal output of PERSONAGE generator during the experimental phases, in which the robot gives information to the human about the compared restaurants in question:

Example 1: The statistics of the generated words and sentences in this example are summarized in Figure (3-8).

- **I**ntroverted Personality: *America is rather excellent. However, Alouette does not provide quite good atmosphere.*
- **E**xtraverted Personality: *Alouette is an expensive bistro (French place in Manhattan), and it offers bad atmosphere and bad stuff. Alva provides nice service, the*

atmosphere is poor though. It is a new American place located near Union Square, you know.

– **A**dapted Personality: *Amarone offers acceptable food, however, the atmosphere is poor. It has friendly waiters, but it is expensive. Although Alva is costly, the food is adequate.*

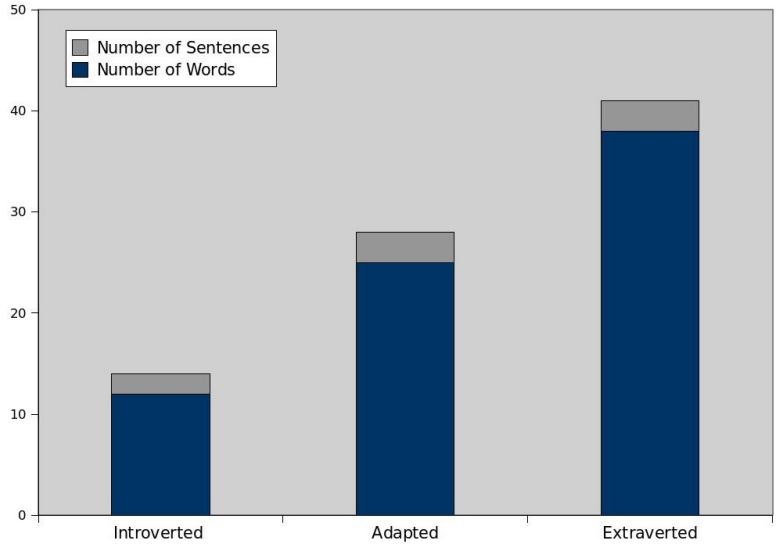


Figure 3-8 – Statistics of the synthesized words and sentences in Example 1, expressed in different personality conditions

Example 2: Similarly to the previous example, the statistics of the synthesized words and sentences are summarized in Figure (3-9).

– **I**ntroverted Personality: *Acappella has nice food with quite outstanding waiters. However, Acacia does not have friendly waiters.*

– **E**xtraverted Personality: *Even if Pho Bang has bad waiters and bad atmosphere, the food is nice. It is a small Vietnamese place in Manhattan. Even if Willow is expensive, the atmosphere is nice, you know. It is a new American place in Milltown. Also, this place offers nice service and nice food.*

– **A**dapted Personality: *Above has adequate waiters, also it offers decent food and pleasant atmosphere. Acacia provides acceptable food and friendly waiters. Its price is 40 USD.*

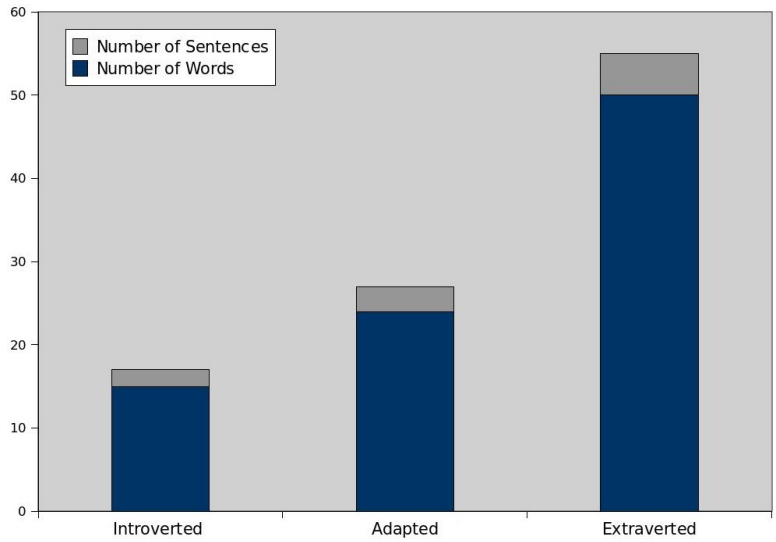


Figure 3-9 – Statistics of the synthesized words and sentences in Example 2, expressed in different personality conditions

Example 3: Finally, the statistics of the synthesized words and sentences are summarized in Figure (3-10).

- **I**ntroverted Personality: *Alfama provides quite good atmosphere and rather outstanding stuff. While, Bar Odeon does not have nasty food.*
- **E**xtraverted Personality: *America has bad stuff and bad atmosphere. Its price is* 27 *USD and it offers poor food. It is a new American place located near Union Square. Bar Odeon provides nice food, even if it is not expensive. Even if this place has rude waiters, the atmosphere is nice, you know. It is a French place in Manhattan.*
- **A**dapted Personality: *Jing Fong's price is* 21 *USD. This place which offers adequate food, is a big Chinese place. Bar Odeon has a pleasant atmosphere. Even if its price is* 44 *USD, the food is acceptable.*

The previous examples reveal the verbal content change of the generated utterances during the experimental phases. The formulation of the generated sentences could be manipulated through the tuning parameters of PERSONAGE generator. This variation made the participants feel that the robot is expressing more details in the extraverted condition than in the introverted condition, also it clarified the verbal content difference between the adapted personality condition and the other personality conditions.

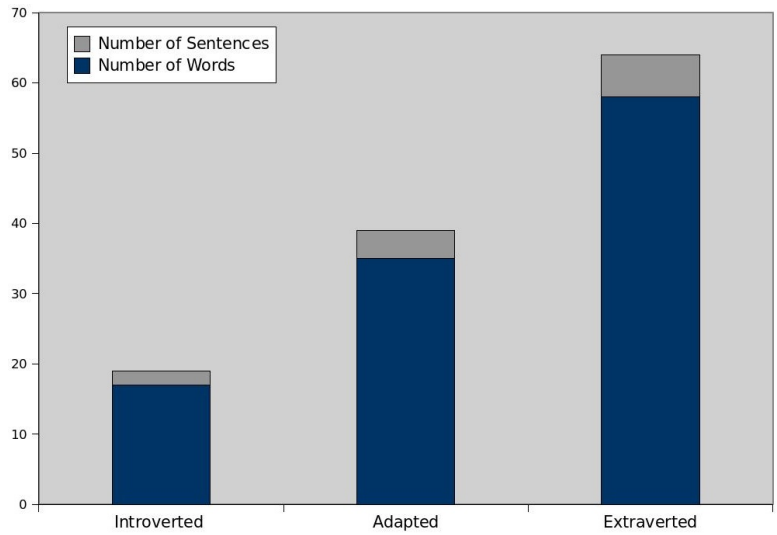


Figure 3-10 – Statistics of the synthesized words and sentences in Example 3, expressed in different personality conditions

The average duration of a single interaction in a given condition was varying between around 3 and 4 minutes. The system was evaluated based on user introspection (i.e., questionnaires). At the end of each experimental phase, each participant completed one questionnaire designed to evaluate and judge: the synchronization between the generated robot's speech and gestures, human's impression about the reflected robot's personality, the interaction with the robot, etc. All questions (i.e., 24 question in total) were presented on a 7-point Likert scale.

3.7 Experimental results

The subject pool consisted of 21 participant (i.e., 7 female and 14 male; 12 introverted and 9 extraverted). Introversion and extraversion are considered belonging to the same personality continuum scale; consequently, having a high score in one of them means having a corresponding complementary low score in the other one. Young [1927] and Jung et al. [1976] proposed a middle group of people in-between introverts and extraverts, called ambiverts, who have both introverted and extraverted features. The ambiversion range on the extraversion-introversion personality scale is equally distributed over the extraversion-ambiversion and ambiversion-introversion intervals. Supposing an ideal ambivert score

is equal to 50%; therefore, we considered the participants with at least 25% introverted functions (i.e., with score less than or equal to 37.5%) to be introverted. Similarly, we considered the participants with at least 25% extraverted functions (i.e., with score greater than or equal to 62.5%) to be extraverted. In this study, all of the calculated personality scores were not included in the considered ambiversion interval (i.e., between 37.5% and 62.5%). Therefore, our analysis focuses only on two categories of participants: introverts and extraverts. The experimental design was based on the within-subjects design. The four experimental phases validating the stated experimental conditions in Section (3.6.2), were randomly ordered. The recruited participants were ENSTA-ParisTech undergraduate and graduate students whose ages were varying between 21-30 years old.

In order to test the first hypothesis, all the participants were exposed to two conditions: introverted robot and extraverted robot. In the introverted robot condition, the generated robot's gestures were narrow, slow, and at a low rate. Contrarily, in the extraverted condition, the generated robot's gestures were broad, quick, and at a high rate (Section 3.5). The content of the generated speech is also based on personality; the robot gave more details in the extraverted condition than in the introverted condition.

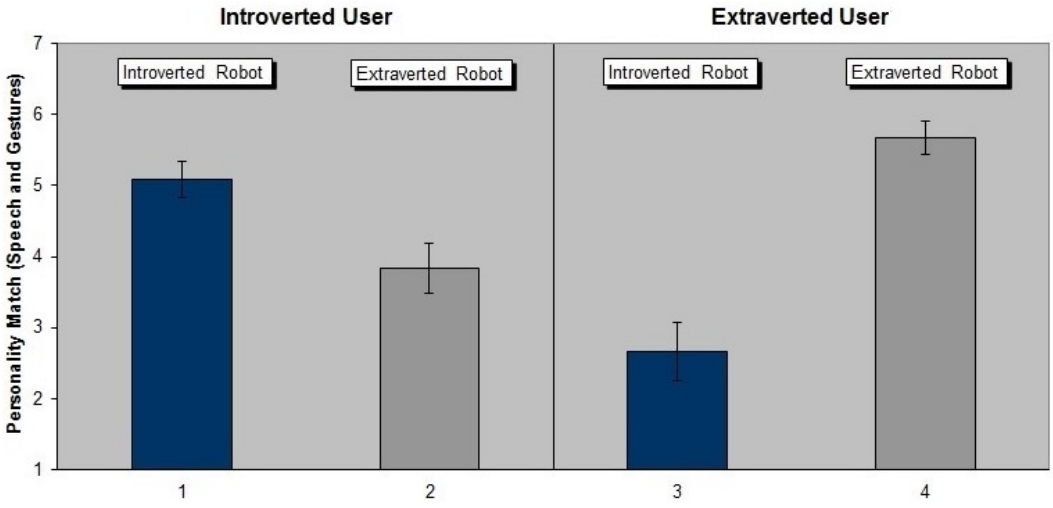


Figure 3-11 – Personality matching for the introverted and extraverted robot conditions

Our ANOVA analysis showed that the extraverted individuals perceived the extraverted robot as significantly more close to their personality than the introverted robot ($F[1,17] = 40.5$, $p < 0.01$). A similar tendency was observed for the introverted individuals who pre-

ferred the introverted robot to the extraverted robot ($F[1,23] = 7.76$, $p = 0.0108$) (Figure 3-11). All the participants (introverted and extraverted together) considered that the robot's speech and gestures were semantically matched (i.e., there was a matching in the content's meaning of both speech and gestures based on the participants' observations) [McNeill, 1992, 2000, 2005; Beattie and Sale, 2012], significantly more in the extraverted condition than in the introverted condition ($F[1,41] = 9.29$, $p = 0.0041$). However, when the user's extraversion-introversion personality trait was included in the analysis, this aspect was significant only for the extraverted individuals ($F[1,17] = 6.87$, $p = 0.0185$).

When the participants were asked about their preferences for the speed of gestures, the extraverted users preferred the extraverted robot with faster movements more than the introverted robot ($F[1,17] = 9.71$, $p = 0.0066$) (Figure 3-12), while the introverted users preferred the introverted robot with slower movements ($F[1,23] = 16.65$, $p = 0.0005$) more than the extraverted robot. These findings are in concordance with Eysenck [1953, 1991] and Eysenck and Eysenck [1968], which linked the extraversion-introversion personality dimension to the activity level, considering the high activity level as an extraverted feature, meanwhile the low activity level tends more to characterize introversion.

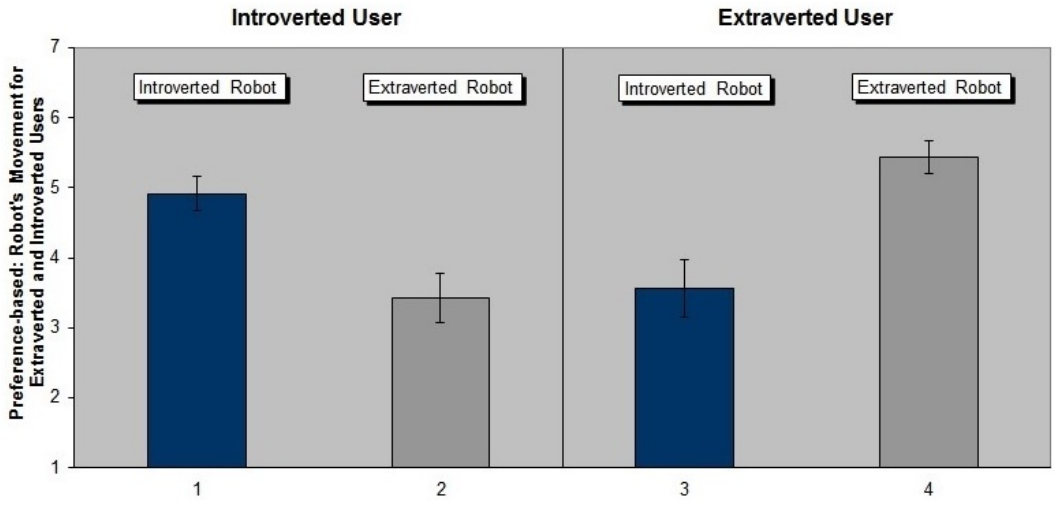


Figure 3-12 – Preference of the introverted and extraverted users for the robot's movement

For the second hypothesis, two other conditions have been examined with all the participants: adapted combined robot's behavior (i.e., gestures and speech are adapted to the user's extraversion-introversion personality trait), and adapted speech-only robot's be-

havior. The participants found the adapted combined robot's behavior more engaging than the adapted speech-only robot's behavior ($F[1,41] = 13.16,\ p = 0.0008$) (Figure 3-13). Through ANOVA test, we found that the adapted speech-only robot's behavior was significantly considered less appropriate ($F[1,41] = 20.16,\ p < 0.01$), and less social ($F[1,41] = 9.137,\ p = 0.004$) than the adapted combined robot's behavior. Moreover, the participants (i.e., the introverted and extraverted participants together) found that the execution of arm movements was fluid with an average score of $M = 4.2$ on a 7-point Likert scale (fluidity is an independent feature of the extraversion-introversion effect on gesture characteristics). At the same time, they agreed that the robot's speech and gestures were semantically matching, and that they were well synchronized with average scores of $M = 5.05$, $SD = 0.59$ and $M = 4.96$, $SD = 0.74$, respectively.

The participants agreed that the combined use of speech and gestures appeared natural with an average score of $M = 4.72$, $SD = 0.95$. On the other hand, when asked if the robot was helpful, no significant difference was found between the adapted speech-only and the adapted combined robot's behaviors with average scores of $M = 5.19$, $SD = 1.36$ and $M = 5.57$, $SD = 1.54$, respectively. The previous results confirm that personality plays an important role in interaction, so that it controls the human's perception and preference for the robot, which makes it an important factor to consider in interaction contexts.

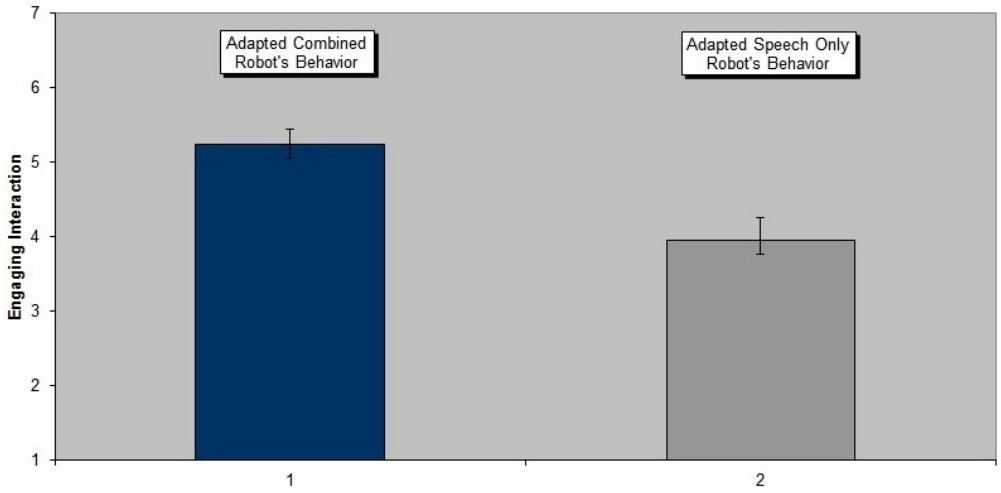


Figure 3-13 – Engaging interaction: adapted combined and adapted speech-only robot's behavior conditions

3.8 Discussion

In this chapter, we investigated the similarity attraction principle within a human-robot interaction scenario, in which the robot adapts its combined multimodal behavior to the interacting human's personality, and we explored the perception of the interacting human for the generated behavior. Moreover, we investigated the effect of the robot's multimodal behavior expressed through speech and gestures on interaction compared to the single-modal behavior expressed through speech only.

The obtained results validated that the behavior of the robot was more preferred when it got adapted to the interacting human's personality. Figure (3-11) illustrates the human's personality-based preference for the robot's behavior, and reveals the binary perception of the extraverted users for the robot's introverted and extraverted conditions. To the contrary, some of the introverted users had a remarkable preference for the extraverted condition of the robot, however this preference was not dominant, so that the similarity attraction principle was validated. This variance in the perception of the robot's behavior between introverts and extraverts reveals the difficulty in setting up clear borders that could separate experimentally between the similarity and complementarity attraction principles. We argue that both of them could co-exist during interaction, however this needs an elaborate study and a big number of participants for validation. On the other hand, the results proved the important role of the robot's multimodal behavior in making the interaction more engaging with the respect to the single-modal behavior (Figure 3-13). This logical result opens the door to other broader studies that employ more communicative cues like facial expressions so as to investigate and compare between the effects of different single and combined modalities of communication on interaction.

3.9 Conclusions

This chapter describes a complete architecture for generating a combined verbal and non-verbal robot's behavior based on the interacting human's personality traits. The personal-

ity dimensions of the interacting human are estimated through a psycholinguistic analysis of speech content. Furthermore, PERSONAGE generator uses the calculated personality scores in order to generate a corresponding text adapted to the interacting human's personality. Afterwards, BEAT toolkit is used in order to generate different kinds of gestures corresponding to the input text (in parallel with our developed general metaphoric gesture generator, which generates gestures based on human's speech).

Our work validates the necessity of having human-robot personality matching for a more appropriate interaction, and shows that an adapted combined robot's behavior through gestures and speech is more engaging and natural than the adapted speech-only robot's behavior. Besides, this chapter proves that extraverts prefer high speed robot's movements contrarily to introverts, and that the perceived semantic matching between the generated speech and gestures by the robot, was higher in the extraverted condition than in the introverted condition. For the future work, we are interested in realizing a more dynamic synchronization between the affiliate and the stroke phase. Besides, we are interested in extending PERSONAGE language generator in order to include other domains than tourism and restaurants. This work has been published in Aly and Tapus [2013a], and is under submission in Aly and Tapus [2015c].

Chapter 4

Prosody-Based Adaptive Head and Arm Metaphoric Gestures Synthesis

In human-human interaction, the modalities of the communication process could be split into: verbal, nonverbal (i.e., gestures), and/or paraverbal (i.e., prosody). The linguistic literature shows that the paraverbal and nonverbal cues are naturally aligned and synchronized, however the natural mechanism of this synchronization is still unexplored. The encountered difficulty during the coordination between voice prosody and head-arm metaphoric gestures concerns the conveyed meaning, the way of performing gestures with respect to the characteristics of prosody, the relative temporal arrangement, and the coordinated organization in the phrasal structure of the utterance. In this chapter, we focus on the mechanism of mapping the prosodic characteristics of speech to head-arm metaphoric gestures in order to generate an adapted robot's behavior to human's emotion. Prosody patterns and the motion curves of head-arm gestures are segmented and aligned separately into parallel Hidden Markov Models (HMM), composing a coupled model (CHMM). A set of head-arm gestures will be synthesized through the CHMM, given an observed audio sequence. An audio-video database covering different emotions was created for validating this study. The obtained results show the effectiveness of the proposed methodology.

4.1 Introduction

Developing intelligent robots capable of behaving and interacting naturally and generating appropriate social behaviors in different interaction contexts in order to make human users believe in the robot's communicative intents, is not a trivial task. The work described in this chapter 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 paraverbal cues. It presents a new methodology that allows the robot adapting automatically its head-arm gestural behavior to the user's emotion, and therefore, to produce a personalized human-robot interaction.

Humans use gestures, postures, and speech in communicative acts. McNeill [1992] and Kendon [1980, 1994] 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 paraverbal information, but also important communicative nonverbal 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 Eyereisen and Lannoy [1991] and Shroder [2009], shows a relationship between speech 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 [Kendon, 1970; Mey, 2001]. 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 and Friesen [1969] identified 5 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), and (5) adaptors (e.g., scratching). On the other hand, Kendon [1982] criticized the classification of Ekman

for neglecting the linguistic phenomena. He proposed a new classification for gestures of 4 categories: (1) gesticulation (e.g., gestures accompanying speech), (2) pantomime (e.g., sequence of gestures with a narrative structure), (3) emblem (e.g., Ok-gesture), and (4) signs (e.g., sign language). McNeill [1992, 2000] collected these four types in a continuum called *Kendon's continuum*. This continuum was later elaborated into 4 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), and (4) beats (e.g., finger movements performed side to side with the rhythmic pulsation of speech).

The important meaning of metaphoric gestures representing abstract ideas has been studied through the conceptual metaphor theory [Lakoff and Johnson, 1980, 1999], which states that linguistic metaphors (e.g., a *long* way, a *high* building, etc.) show that a lot of our abstract ideas in mind, which have no spatial representation in the world, are being expressed into physical space through metaphoric gestures. The psychological experiments conducted in Cienki [2000] and Casasanto and Lozano [2007] stated that metaphoric gestures provide the speaker with an important internal cognitive function, like facilitating access to some appropriate spatial words or concepts so as to make the meaning of the utterance understood easily by the listener. Similarly, these experiments shown that the absence of metaphoric gestures may increase stuttering during speech with spatial content. The important role of metaphoric gestures in human-human communication, increases the necessity of making the robot capable of generating similar gestures that can represent abstract ideas as naturally as humans do.

On the other hand, iconic and metaphoric gestures (according to McNeill's categorization) constitute the main body of the generated nonverbal 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. [2001] 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) [Cassell et al., 2000, 2001], and humanoid robots [Aly and Tapus, 2013a]. This system can synthesize various categories of gestures (including iconic gestures) except for metaphoric gestures. Similarly, Pelachaud [2005] developed the 3D virtual conversational agent GRETA, which can generate a synchronized multimodal behavior to the human users. GRETA can generate all kinds of gestures regardless of the domain of interaction, unlike the other 3D conversational agents (e.g., the conversational agent MAX) [Kopp and Wachsmuth, 2004; Kopp et al., 2008]. 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) [DeCarolis et al., 2004], which is used as a script language to control the animation of the agent. Recently, an interesting architecture has been discussed in Le et al. [2012], 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 Neff et al. [2008]; Kipp et al. [2007], in which they proposed a probabilistic synthesis method trained on hand-annotated videos. Similarly, another system was illustrated in Ng-Thow-Hing et al. [2010], 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. Generally, 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. Besides, it makes them unable to measure emotions that influence body language, which may hinder generating a robot's behavior adapted to human's emotion [Jensen et al., 2000].

Another interesting approach towards generating iconic gestures was discussed in Kopp and Wachsmuth [2004]; Kopp et al. [2008]. They developed the 3D virtual conversational agent MAX, which uses synchronized speech and gestures in order to interact multimodally with humans (e.g., describing a place multimodally based on some prescribed dimensional knowledge about that place). This approach 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). Bennewitz et al. [2007]

developed the communication museum tour-guide robot FRITZ, which can perform a specific set of gestures (e.g., greeting gesture, come-closer gesture, disappointment gesture, head gestures (e.g., nods to agree or disagree), and pointing gestures) accompanied by speech. However, it was not able to generate any continuous metaphoric or iconic gestures. Generally, the aforementioned approaches are -still- away from considering human's emotion when generating a multimodal robot's behavior, in which voice prosody correlates with human's emotion, in addition to the dynamic characteristics of the human body language.

On the way towards generating an animation script based on speech features, Bregler et al. [1997] and Brand [1999] studied the relationship between phonemes and facial expressions. Sargin et al. [2008] proposed a time-costly probabilistic model to synthesize metaphoric head gestures from voice prosody. A similar approach was discussed in Deng et al. [2004], which uses the features of head gestures and voice prosody in order to create a training database for a statistical model that can generate a set of motion sequences for the 3D agents. Another interesting approach was discussed in Levine et al. [2009, 2010], which selects some segments from a motion database based on an audio input, and then synthesizes these segments into head-arm metaphoric gestures that could animate the 3D agents. Despite these interesting approaches, the relationship between human's emotion and head-arm gestures is still incompletely addressed, which constituted our motivation for this work.

The rest of the chapter is structured as following: Section (4.2) presents an overview of the system architecture, Section (4.3) presents the database used in this work, Section (4.4) illustrates the analysis of gesture kinematics, Section (4.5) illustrates data segmentation, Section (4.6) validates the chosen speech and gestures characteristics, Section (4.7) describes data quantization, Section (4.8) explains the coupling between speech and head-arm gestures using the CHMM, Section (4.9) describes and validates the synthesis of customized head-arm gestures to human's emotion, and finally, Section (4.10) concludes the chapter.

4.2 System architecture

The system is coordinated through three stages, as illustrated in Figure (4-1). 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 (i.e., the pitch-intensity curves of speech and the motion curves of gesture). Afterwards, the extracted characteristic curves of speech and gestures go to the segmentation phase and then to the Coupled Hidden Markov Models (CHMM) phase. The CHMM is composed of a multi-stream collection of parallel HMM [Rabiner, 1989; Rezek et al., 2000; Rezek and Roberts, 2000], through which new *adapted* head-arm gestures will be synthesized (i.e., Stage 2) based on the prosodic cues of a new speech-test signal, which will undergo the same phases of the training stage. The main advantages of using the CHMM for generating gestures, are: (1) the random variations of the generated gestures, which make them more human-like than when a fixed gesture dictionary is used, and (2) the ability to generate gestures of varying duration and amplitude adapted to human's prosodic cues.

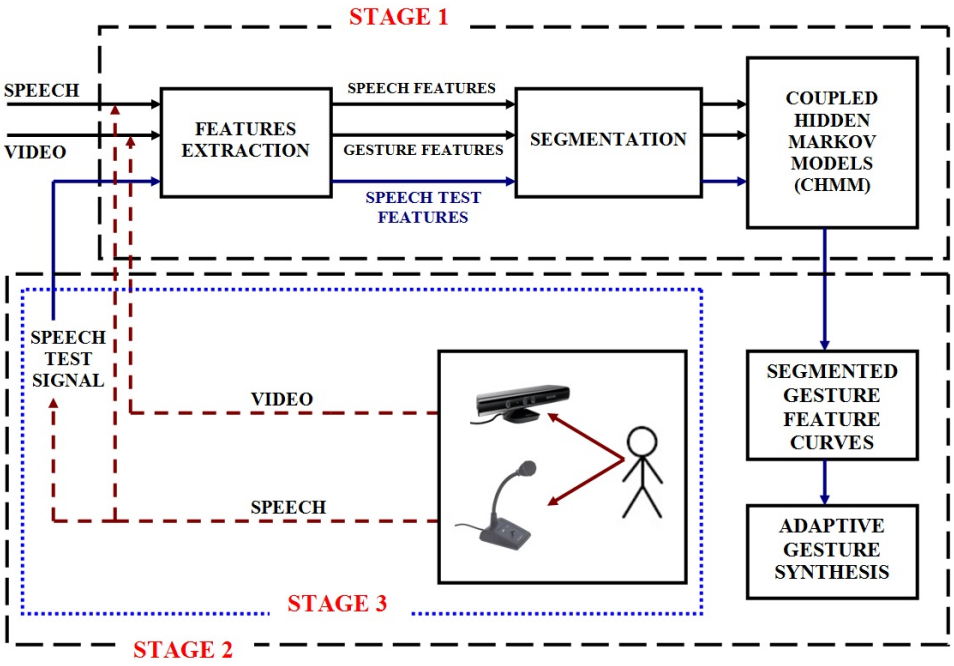


Figure 4-1 – Metaphoric gesture generator architecture

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 its surroundings. This requires the Kinect sensor, which can calculate in real time the rotation curves of the head and arms articulations, and a microphone in order to receive human's speech. Afterwards, both the captured audio and

video data will follow the previously explained phases of the training Stage 1 so as to increase the robot's ability to generate more appropriate gestures. Similarly, when a new speech-test signal from one of the individuals around the robot is present, it will follow the phases of the test Stage 2. In this work, we focus on Stages $1-2$ and we validate their theoretical bases. However, Stage 3 represents a future experimental stage towards creating a successful multimodal human-robot interaction architecture.

4.3 Database

The synchronized audio-video database used in this work 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 7 categories of pure continuous emotion expression: sadness, disgust, surprise, happiness, anger, fear, and neutral. The chosen emotions constitute the main primary emotions stated by most of the contemporary theories of emotion [Ekman et al., 1982; Plutchik, 1991]. We have not tried to include any complex emotion [Plutchik, 1991] 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`. This database extends the neutral emotion database created by Levine et al. [2009] and respects the constraints of data acquisition and processing.

4.4 Gesture kinematic analysis

The hierarchical construction of the 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*) [Aggarwal and Cai, 1999]. This parent-child relationship of body segments allows the inheritance of motion characteristics from

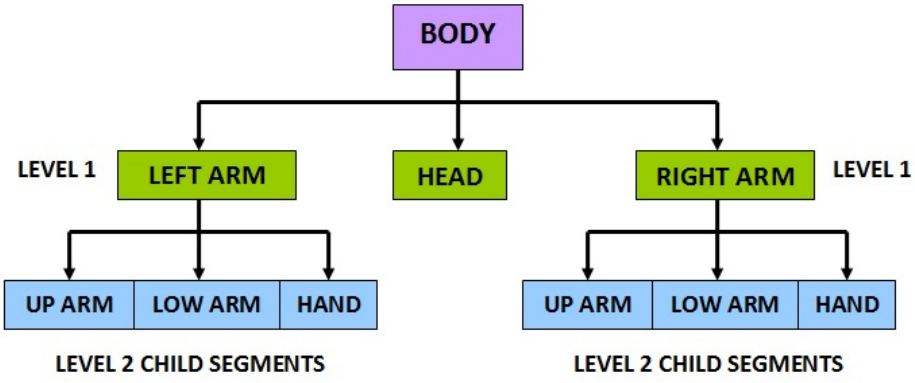


Figure 4-2 – Parent-Child hierarchy

the parent to child segments, and vice versa. In this chapter, we assume that the legs, waist, and torso segments will keep static during emotion expression, so that for the parent segment *body*, the child segments will be limited to: *head*, *left arm*, and *right arm*, as illustrated in Figure (4-2). 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 would be characterized only in terms of the linear velocity and acceleration).

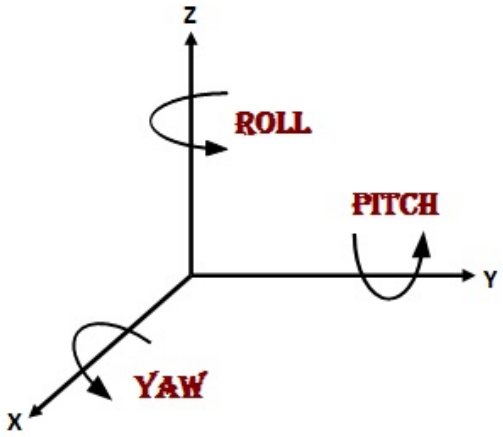


Figure 4-3 – Roll-Pitch-Yaw rotations

4.4.1 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 coor-

dinate axes indicated in Figure (4-3), the rotation about the reference z-axis is denoted by ϕ (Roll), 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 (Equation 4.1) [Ang and Tourassis, 1987]:

$$\omega = \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = \begin{pmatrix} 0 & -sin\phi & cos\phi\ cos\theta \\ 0 & cos\phi & sin\phi\ cos\theta \\ 1 & 0 & -sin\theta \end{pmatrix} \begin{pmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{pmatrix} \tag{4.1}$$

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 (Equation 4.2):

$$\begin{aligned} \dot{\omega} = \begin{pmatrix} \dot{\omega}_x \\ \dot{\omega}_y \\ \dot{\omega}_z \end{pmatrix} &= \begin{pmatrix} 0 & -sin\phi & cos\phi\ cos\theta \\ 0 & cos\phi & sin\phi\ cos\theta \\ 1 & 0 & -sin\theta \end{pmatrix} \begin{pmatrix} \ddot{\phi} \\ \ddot{\theta} \\ \ddot{\psi} \end{pmatrix} \\ + \begin{pmatrix} -cos\phi & -sin\phi\ cos\theta & -cos\phi\ sin\theta \\ -sin\phi & cos\phi\ cos\theta & -sin\phi\ sin\theta \\ 0 & 0 & -cos\theta \end{pmatrix} &\begin{pmatrix} \dot{\phi}\ \dot{\theta} \\ \dot{\phi}\ \dot{\psi} \\ \dot{\theta}\ \dot{\psi} \end{pmatrix} \end{aligned} \tag{4.2}$$

4.4.2 Body Segment Parameters Calculation

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

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

The literature of kinetics illustrates big efforts towards stating a unified mathematical representation of the human body including the previously mentioned parameters, however the outcome was always approximate and different from a research to another [Zatsiorsky and Seluyanov, 1979; Plagenhoef et al., 1983; Leva, 1996]. For the calculation of the mass of each body segment required for gesture segmentation (Section 4.5.1), we used the highly cited relationships stated in Kroemer et al. [1994], as indicated in Equation (4.3) (where M denotes the total body mass):

$$\begin{aligned} Head\ Mass &= 0.0307 * M + 2.46 \\ Up\ Arm\ Mass &= 0.0274 * M - 0.01 \\ Low\ Arm\ Mass &= 0.70 * (0.0233 * M - 0.01) \\ Hand\ Mass &= 0.15 * (0.0233 * M - 0.01) \end{aligned} \tag{4.3}$$

Similarly, the length of each body segment could be calculated in terms of the person's height using the following approximate relationships (Equation 4.4) [Winter, 2009]:

$$\begin{aligned} Neck\ Length &= 0.052 * Person\ Height \\ Up\ Arm\ Length &= 0.187 * Person\ Height \\ Low\ Arm\ Length &= 0.1455 * Person\ Height \\ Hand\ Length &= 0.108 * Person\ Height \\ Shoulder\ Length &= 0.129 * Person\ Height \end{aligned} \tag{4.4}$$

The height and mass parameters of human are only required for constructing the initial learning database of the CHMM, however they will not be required during an online human-robot interaction, in which the robot will use the previously trained CHMM for synthesizing head-arm metaphoric gestures.

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 4.5), in addition to calculating the Denavit-Hartenberg parameters of the head (Appendix B). Meanwhile, the length of the shoulder is required for calculating the forward kinematics model of the arm (Section 4.4.3).

The proximal distance from the center of mass (CM) of each segment to the nearest articulation could be calculated in terms of the length of the segment, as illustrated in Equation (4.5) (where the left and right arm segments are symmetric and have equal lengths) [Plagenhoef et al., 1983]:

$$\begin{aligned} d_{CMHead \rightarrowtail UpNeck} &= Neck\ Length \\ d_{CMUpArm \rightarrowtail Shoulder} &= 0.447 * Up\ Arm\ Length \\ d_{CMLowArm \rightarrowtail Elbow} &= 0.432 * Low\ Arm\ Length \\ d_{CMHand \rightarrowtail Wrist} &= 0.468 * Hand\ Length \end{aligned} \tag{4.5}$$

From Equations (4.1), (4.2), and (4.5), the linear velocity and acceleration of body segments could be formulated as following (Equations 4.6 and 4.7):

$$\begin{pmatrix} V_{Head} \\ V_{UpArm} \\ V_{LowArm} \\ V_{Hand} \end{pmatrix} = \begin{pmatrix} \omega_{Head} * d_{CMHead \rightarrowtail UpNeck} \\ \omega_{UpArm} * d_{CMUpArm \rightarrowtail Shoulder} \\ \omega_{LowArm} * d_{CMLowArm \rightarrowtail Elbow} \\ \omega_{Hand} * d_{CMHand \rightarrowtail Wrist} \end{pmatrix} \tag{4.6}$$

$$\begin{pmatrix} A_{Head} \\ A_{UpArm} \\ A_{LowArm} \\ A_{Hand} \end{pmatrix} = \begin{pmatrix} \dot{\omega}_{Head} * d_{CMHead \rightarrowtail UpNeck} \\ \dot{\omega}_{UpArm} * d_{CMUpArm \rightarrowtail Shoulder} \\ \dot{\omega}_{LowArm} * d_{CMLowArm \rightarrowtail Elbow} \\ \dot{\omega}_{Hand} * d_{CMHand \rightarrowtail Wrist} \end{pmatrix} \tag{4.7}$$

4.4.3 Forward Kinematics Model of the Arm

The 3 articulations of the human arm have 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 (which is concerned with calculating the position of the end-effector) through the 7 DOF of the human arm by a series of homogeneous transformation matrices [Asfour and Dillmann, 2003]. The transformation matrix required to transform the coordinate frame *i-1* to *i* is illustrated in Equation (4.8) (where $C\theta$ denotes $Cos(\theta)$ and $S\theta$ denotes $Sin(\theta)$):

$T_{i-1\rightarrowtail i}$	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 4.1 – Denavit-Hartenberg parameters of the left and right arms

$$T_{i-1\rightarrowtail i} = \begin{pmatrix} C\theta_i & -C\alpha_i S\theta_i & S\alpha_i S\theta_i & a_i C\theta_i \\ S\theta_i & C\alpha_i C\theta_i & -S\alpha_i C\theta_i & a_i S\theta_i \\ 0 & S\alpha_i & C\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{4.8}$$

The parameters of the transformation matrix for the arms are defined in Table (4.1). The *highlighted* elements represent the position coordinates (x,y,z) of the joint. Therefore, the position and orientation of arms articulations could be calculated from Equation (4.9):

$$\begin{pmatrix} Position_{\ Shoulder} \\ Position_{\ Elbow} \\ Position_{\ Wrist\ (End\ Effector)} \end{pmatrix} = \begin{pmatrix} \prod_{i=1}^{3} T_i \\ \prod_{i=1}^{4} T_i \\ \prod_{i=1}^{7} T_i \end{pmatrix} \tag{4.9}$$

Finally, the displacement of arms articulations could be calculated directly from the Euclidean distance between the position coordinates of each articulation in frames i and $i+1$ of the video data.

4.5 Multimodal data segmentation

The segmented speech and gestures are modeled separately on different Hidden Markov Models (HMM), which compose the coupled audio and video chains of the CHMM (Section 4.8). Figure (4-4) illustrates the structure of the parallel HMM. It is composed of N parallel states, where each contains M observations (the number of observations could be different in the audio and video chains). The goal of the transition between the states S_{END}

and S_{START}, is to continue the transitions between the states of the HMM model (from State 1 to State N) in a sequential way. Each state of the video chain represents a complete gesture, while each state of the audio chain represents the corresponding audio segment (i.e., syllable) to the segmented gesture. Therefore, gestures will be segmented first using the algorithm discussed below, then the boundaries of the corresponding audio segments will be calculated in terms of gesture boundaries.

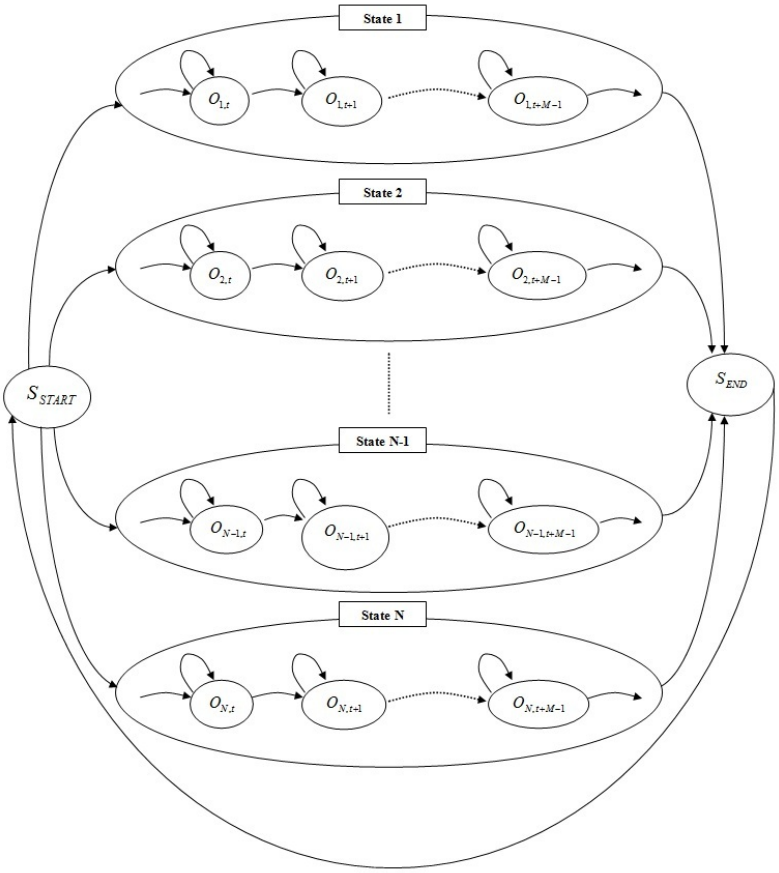


Figure 4-4 – Hidden Markov Model (HMM) structure

4.5.1 Gesture Segmentation

The difficulty behind gesture segmentation lies in the fact that people perceive gesture boundaries in different manners within a continuous motion sequence [Badler et al., 2000; Kahol et al., 2003], which poses a potential challenge towards defining unified characteristics for gesture segmentation. The literature reveals 2 main techniques for gesture segmentation: (1) pose-based segmentation, which is inappropriate for segmenting metaphoric

gestures from a continuous gesture sequence [Bobick and Wilson, 1995; Campbell and Bobick, 1995], and (2) low-level descriptors based segmentation (e.g., velocity and acceleration), which is used in this study [Goddard, 1994; Lee and Xu, 1996; Wang et al., 2001]. Velocity and acceleration based techniques consider each local minimum point as a gesture boundary, which is not totally a valid assumption, because not all the local minimum points of velocity or acceleration curves represent real gesture boundaries [Kahol et al., 2003]. Consequently, other velocity and acceleration based descriptors (that can better characterize the activity of a body segment): *force (F), momentum (M), and kinetic energy (KE)*, are used for gesture segmentation. Equation (4.10) indicates the mathematical formulas for calculating the activity of body segments, in terms of the mass, the velocity, and the acceleration obtained from Equations (4.3), (4.6), and (4.7):

$$\begin{aligned} F_{Segment} &= Mass_{Segment} * A_{Segment} \\ M_{Segment} &= Mass_{Segment} * V_{Segment} \\ KE_{Segment} &= \frac{1}{2} * Mass_{Segment} * V_{Segment}^2 \end{aligned} \tag{4.10}$$

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.

4.5.2 Speech Segmentation

After calculating gesture boundaries, the corresponding boundaries of audio segments could be simply derived as in Equation (4.11) (where *A* denotes *Audio*, *G* denotes *Gesture*, and *Fs* denotes the audio *Sampling Frequency*):

$$A_{Boundaries} = G_{Boundaries} * Frame\,Time * F_S \tag{4.11}$$

These audio segments (i.e., syllables) and the accompanying gestures constitute the speech-gesture multimodal database that is used in training the CHMM, as will be explained in details later on.

4.6 Multimodal data characteristics validation

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

4.6.1 Body Gestural Behavior Recognition in Different Emotional States

Each gesture performed by a body segment is characterized in terms of the linear velocity, the linear acceleration, the position, and the displacement of the segment. Afterwards, common statistic measurements: *mean, variance, maximum, minimum, and range* have been calculated for the characteristic motion curves, composing both the learning and test databases. Data was cross validated using the Support Vector Machine (SVM) algorithm. Table (4.2) illustrates the obtained recognition scores of the total body gestural behavior in different emotional states, which validates the pertinence of the chosen dynamic characteristics for gesture recognition.

Emotion	Body Gestural Behavior
Sadness	85.4%
Disgust	79.3%
Surprise	89.2%
Happiness	91.1%
Anger	93.9%
Fear	76.3%
Neutral	88.9%

Table 4.2 – Recognition scores of the body gestural behavior in different emotional states based on the dynamic characteristics of gesture

4.6.2 Emotion Recognition Based on Audio Characteristics

Emotion recognition based on the prosodic cues of speech has been the focus of a lot of researches in the literature. Table (4.3) demonstrates the recognition scores of different emotions, which *we have obtained in Chapter (2)* using 3 well-known databases (i.e., GES, GVEESS, and SES) [Aly and Tapus, 2012e]. Meanwhile, the last column indicates the recognition scores of the same emotions using our new database, which is composed of the segmented audio data accompanying the body behaviors under study. These results validate the pertinence of the chosen prosodic characteristics for emotion recognition.

Emotion	GES	GVEESS	SES	NEW DATABASE
Sadness	86.9%	90.1%	94.1%	95.3%
Disgust	92.1%	91.7%	-	75.2%
Surprise	-	-	95.7%	81.4%
Happiness	86.9%	88.5%	75.1%	84.6%
Anger	80.8%	88.7%	79.8%	96.9%
Fear	-	85.7%	-	82.3%
Neutral	83.7%	-	89.5%	91.4%

Table 4.3 – Recognition scores of different emotions based on the prosodic cues of speech. Empty spaces represent the not-included emotions in the databases (Table 2.1).

4.7 Data quantization

Speech and gestures characteristic curves are quantized before training the CHMM. The common inflection points (i.e., the points at which the curve's curvature changes sign from positive to negative or from negative to positive) of the pitch and intensity curves are calcu-

lated, afterwards the resulting segmented trajectories of both curves are labeled as indicated in Table (4.4). Similarly, the common inflection points of gesture motion curves are calculated and the corresponding trajectory labels are attributed as indicated in Table (4.5), 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 4.4 – 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 4.5 – Gesture segmentation labels (*D denotes Displacement, V denotes Velocity, A denotes Acceleration, and P denotes Position*)

4.8 Speech to gesture coupling

A typical CHMM structure is shown in Figure (4-5), where the circles represent the discrete hidden nodes/states. Meanwhile, the rectangles represent the observable nodes/states,

which contain the observation sequences of the characteristics of speech and gestures. According to the sequential nature of speech and gestures, the CHMM structure is of type lag-1, in which the 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 (λ_C) could be defined by the following parameters stated in Equation (4.12):

$$\begin{aligned} \pi &= \{\pi_0^C(i)\} = P(q_1^C = S_i) \\ A &= \{a_{i|j,k}^C\} = P(q_t^C = S_i | q_{t-1}^{audio} = S_j, q_{t-1}^{video} = S_k) \\ B &= \{b_t^C(i)\} = P(O_t^C | q_t^C = S_i) \end{aligned} \tag{4.12}$$

where π denotes the initial state probability, A denotes the transition probability, B denotes the observation probability, $S_{i,j,k}$ denote different states of the model, $C \in \{audio, video\}$ denotes the audio and video chains respectively, and q_t^C denotes the state of the coupling node in the C_{th} stream at time t [Nean et al., 2002].

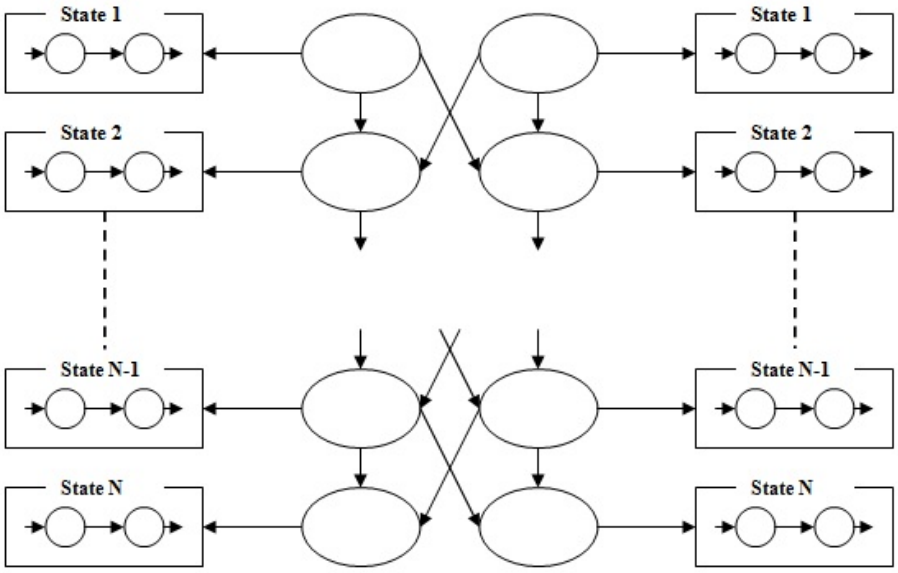


Figure 4-5 – Coupled Hidden Markov Models (CHMM) structure

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 sequence of the observable states in the audio chain, $B_{1..N} = \{b_1, \cdots, b_N\}$ is the sequence of the observable states in the video chain, and $S = \{X_{1..N}, Y_{1..N}\}$ is the sequence of states of the audio and video chains

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 in Equation (4.13) [Rezek and Roberts, 2000]:

$$f(\lambda_C) = P(X_1)\ P(Y_1) \prod_{t=1}^{T} P(A_t|X_t)\ P(B_t|Y_t)\ P(X_{t+1}|X_t,Y_t)\ P(Y_{t+1}|X_t,Y_t),\ 1 \leq T \leq N \quad (4.13)$$

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, which both have a multivariate Gaussian distribution.
- $P(X_{t+1}|X_t,Y_t)$ and $P(Y_{t+1}|X_t,Y_t)$ are the state transition probabilities of the audio and video chains.

The training of the CHMM differs from the standard HMM in the expectation step, while they are both identical in the maximization step, which tries to maximize Equation (4.13) in terms of the expected parameters. The expectation step of the CHMM could be defined in terms of the forward and backward recursions. For the forward recursion, we define a variable α for each of the audio and video chains: α_t^{audio} and α_t^{video} at $t = 1$, as indicated in Equation (4.14):

$$\begin{aligned} \alpha_{t=1}^{audio} &= P(A_1|X_1)P(X_1) \\ \alpha_{t=1}^{video} &= P(B_1|Y_1)P(Y_1) \end{aligned} \quad (4.14)$$

Afterwards, the variable α will be calculated incrementally at any arbitrary moment t, leading to the following final Equation (4.15):

$$\begin{aligned} \alpha_{t+1}^{audio} &= P(A_{t+1}|X_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video} P(X_{t+1}|X_t,Y_t)\ dX_t\ dY_t \\ \alpha_{t+1}^{video} &= P(B_{t+1}|Y_{t+1}) \int\int \alpha_t^{audio}\alpha_t^{video} P(Y_{t+1}|X_t,Y_t)\ dX_t\ dY_t \end{aligned} \quad (4.15)$$

Meanwhile, for the backward direction, there will not be any split in the calculated recursions, which could be expressed as following (Equation 4.16):

$$\beta_{t+1}^{audio,video} = P(O_{t+1}^{N}|S_t) = \int\int P(A_{t+1}^{N}, B_{t+1}^{N}|X_{t+1}, Y_{t+1})\, P(X_{t+1}, Y_{t+1}|X_t, Y_t) \; dX_{t+1}\, dY_{t+1} \tag{4.16}$$

4.9 Gesture synthesis validation and discussion

Viterbi decoding algorithm of the CHMM [Rabiner, 1989; Rezek et al., 2000] is concerned with finding the most likely path of observations through the gesture chain of the trained CHMM, given an observed audio sequence. 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 4.5) (in case a specific trajectory class label is repeating within the resulting label sequence of the quantized segments, the mean value of the corresponding segments of *each* motion curve separately, will be calculated). These defined segments of the motion curves will be used after the Viterbi decoding of the CHMM in constructing the synthesized motion curves of gesture by substituting each decoded trajectory class label with its *approximate* corresponding segments of the motion curves.

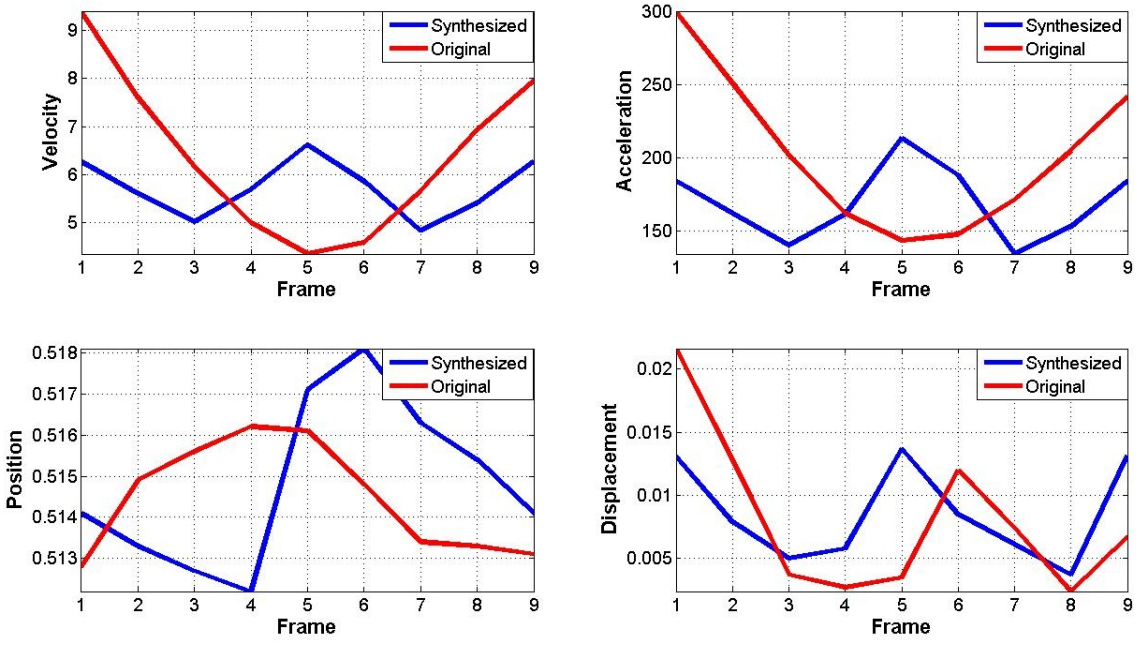


Figure 4-6 – Synthesized motion curves (velocity, acceleration, position and displacement) of a right-arm shoulder's gesture, expressing the emotional state "disgust"

Figure (4-6) illustrates the synthesized motion curves of a shoulder gesture. The first two graphs (i.e., the 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, considering that metaphoric gestures represent abstract ideas. On the other hand, there will not be a big difference between the original and synthesized curves shown in Figure (4-6) in case they get characterized in terms of the statistic measurements required for the classification system explained in Section (4.6.1). This explains the relatively small differences between the obtained recognition scores in Tables (4.2) and (4.6). Table (4.6) 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 the methodology of synthesizing metaphoric gestures discussed in this chapter.

Having calculated the synthesized motion curves of gesture, it is possible to calculate the corresponding rotation angles of arms articulations in order to be modeled on the robot, using the generated position curves, the orientation, and the inverse kinematics model of the arm, as explained in Appendix (A). Similarly, the rotation angles of the head could be calculated in terms of the orientation and the inverse kinematics model of the head, as explained in Appendix (B). 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). A video showing NAO robot generating a sequence of head gestures is available at: `http://perso.ensta-paristech.fr/~tapus/eng/media.html`.

Emotion	Generated Body Gestural Behavior
Sadness	82.3%
Disgust	75.2%
Surprise	81.3%
Happiness	83.5%
Anger	85.6%
Fear	72.4%
Neutral	78.1%

Table 4.6 – Recognition scores of the body gestural behavior generated by the CHMM in different emotional states

An experimental study for validating the synthesized head-arm metaphoric gestures in different emotional states based on human's evaluation, is set in Chapter (5). In this study, the humanoid ALICE robot was used within a narrative human-robot emotional interaction. The emotional content of the synthesized multimodal behavior modeled on the robot, was validated by the human user.

4.10 Conclusions

This chapter discusses how to synthesize adapted head-arm metaphoric gestures to human's speech using the Coupled Hidden Markov Models (CHMM), which is composed of two chains for modeling speech and gestures. The motion curves of gesture are calculated from the rotations of articulations using the dynamic parameters and the kinetic analysis of the human body. These motion curves are segmented by calculating the force, momentum, and kinetic energy of body segments (which are considered as high level descriptors for the hierarchical construction of the human body, that presents it as a connection between the main parent segment "body" and a series of child segments "e.g., left arm", and its dynamic activity), in addition to calculating the total force of the body. The intersection between these descriptors represents the boundary points of gestures in each body segment. Additionally, the pitch-intensity curves of speech are segmented in terms of the calculated boundary points of the segmented gestures. The segmented speech and gesture patterns are used in training the CHMM model, which will synthesize a set of gestures based on the prosodic cues of speech that correlate with human's emotion.

The kinetic analysis of the human body discussed in this chapter uses the mass and height of human in order to construct the offline training database of the CHMM, which could *not* be considered as a limitation in this work. However, when an online human-robot interaction starts, this information about the interacting human will *not* be required, so that adapted gestures will be synthesized through the trained CHMM, given an audio signal.

The obtained recognition scores of speech and gestures, in addition to the synthesized gestures by the CHMM in different emotional states, prove the pertinence of the chosen

speech-gesture characteristics in our analysis. For the future work, we are interested in making the robot able to increase online its initial learning database by acquiring more audio and video data from the nearby humans in the robot's environment, through Stage 3 of the generator architecture (Figure 4-1). This work has been published in Aly and Tapus [2013b], and extends the proposed methodologies of our previous researches [Aly and Tapus, 2011b, 2012b,d].

Chapter 5

Multimodal Adapted Robot's Behavior Synthesis within a Narrative Human-Robot Interaction

In human-human interaction, three modalities of communication (i.e., verbal, nonverbal, and paraverbal) are naturally coordinated so as to enhance the meaning of the conveyed message. In this chapter, we try to create a multimodal coordination between these modalities of communication in order to make the robot behave as naturally as humans do. The proposed system uses videos to elicit certain emotions in human, upon which interactive narratives will start (i.e., interactive discussions between the participant and the robot about each video's content). During each interaction, the robot engages and generates an adapted multimodal behavior (i.e., using facial expressions, head-arm metaphoric gestures, and/or speech) to the emotional content of the video. This synthesized multimodal robot's behavior is evaluated by the interacting human at the end of each emotion-eliciting experiment in order to explore the important effect of these modalities of communication on interaction. Mary-TTS (text to speech toolkit) is employed in order to generate emotional speech, which is used - in parallel - to synthesize adapted head-arm metaphoric gestures [Aly and Tapus, 2013b] (which is considered as a practical validation for the conducted study in Chapter 4). Experiments with ALICE robot are reported.

5.1 Introduction

The need for an intelligent robot that can adapt the emotional content of its synthesized multimodal behavior to the context of interaction so as to increase the credibility of its communicative intents, is increasing rapidly. The literature reveals a lot of elaborate studies that discuss the relationship between emotion from one side, and both the prosodic characteristics of speech and the dynamic characteristics of gestures and facial expressions from the other side.

Gestures, facial expressions, and speech are used together to convey coordinated and synchronized verbal and paraverbal information [Eyereisen and Lannoy, 1991; Shroder, 2009], in addition to some important communicative nonverbal cues that could enhance and complement the conversation, such as: facial expressions, arm-hand movements, and/or head movements. The importance of facial expressions during interaction lies in their ability to clarify the meaning of speech when the signal is deteriorated, in addition to the fact that they can replace or accompany words in a synchronized manner [Ekman, 1979].

The correlation between emotion and speech had been investigated in many researches [Cowie and Cornelius, 2003; Scherer, 2003]. Speech prosody can reflect human's emotion through changes in basic cues, such as: pitch, intensity, rate, and pauses [Cowie and Cornelius, 2003]. The variation in the characteristics of voice prosody for different emotions: anger, disgust, fear, and sadness, was studied in Bachorowski [1999] and Sauter et al. [2010]. Pell and Kotz [2011] studied the process of emotion perception and decoding, in addition to the required time to recognize different emotions based on their prosodic cues. Aly and Tapus [2012e, 2015b] considered the evolutionary nature of emotion [Plutchik, 1991], while studying the cognitive perception of emotions through a fuzzy model.

On the way towards synthesizing emotional speech that can add more naturalness to human-robot and human-computer interactions, the first emotional speech synthesis system was developed based on the rule-based formant synthesis technique [Cahn, 1990b; Murray and Arnott, 1995], but the quality was a bit poor. Another interesting approach based on the diphone concatenation technique that achieved some limited success in expressing spe-

cific emotions, was discussed in Vroomen et al. [1993] and Edgington [1997]. This last technique was later developed to the unit selection technique that tries to avoid interference with the recorded voice during synthesis so as to obtain a better quality, and reported some small success in expressing only three emotions: happiness, anger, and sadness [Iida and Campbell, 2003]. Generally, the previously discussed techniques are missing some explicit control on the prosodic parameters of speech so as to be able to express emotions on a wider scope. This constraint and the quality of the synthesized voice, constituted our inspiration for using a more controllable and efficient text-to-speech engine, like Mary-TTS [Schroder and Trouvain, 2003] in our work.

On the other hand, the basic definition for gesture was given by Kendon [1980] and McNeill [1992]. They defined a gesture as a synchronized body movement with speech, which is related parallelly or complementarily to the meaning of the utterance. The first step towards categorizing gestures, was discussed in Ekman and Friesen [1969]. They proposed 5 gesture categories: (1) emblems (e.g., waving), (2) illustrators (e.g., pointing), (3) facial expressions, (4) regulators (e.g., arm-hand movements and body postures), and (5) adaptors (e.g., scratching). However, Kendon [1983] criticized the proposed gesture categories of Ekman for ignoring the linguistic phenomena. Therefore, he proposed a new classification for gestures of 4 categories: (1) gesticulation (e.g., gestures accompanying speech), (2) pantomime (e.g., sequence of gestures with a narrative structure), (3) emblem (e.g., Ok), and (4) signs (e.g., sign language). McNeill [1992, 2000] presented a more elaborate widely used gesture classification of 4 categories: (1) iconics (e.g., gestures representing images of concrete entities and/or actions), (2) metaphorics (e.g., gestures representing abstract ideas), (3) deictics (e.g., pointing), and (4) beats (e.g., finger movements performed side to side with the rhythmic pulsation of speech).

Several researches in the fields of human-robot interaction and human computer interaction, have focused on synthesizing iconic and metaphoric gestures, which form together (according to McNeill) the major part of the generated nonverbal behavior during human-human interaction. Pelachaud [2005] developed the 3D agent GRETA, which can synthesize a multimodal synchronized behavior to the human users based on an input text. Generally,

GRETA can synthesize gestures of all categories regardless of the domain of interaction, to the contrary of other 3D conversational agents (e.g., the conversational agent MAX) [Kopp and Wachsmuth, 2004; Kopp et al., 2008]. An interesting framework was discussed in Le et al. [2012], which can synthesize a multimodal synchronized behavior for both the 3D agent GRETA and the robot. Cassell et al. [2001] presented BEAT toolkit, which is a rule-based gesture generator. It applies the natural language processing (NLP) algorithms on an input text in order to produce an animation script that can animate both of humanoid robots [Aly and Tapus, 2013a], and virtual agents (e.g., REA agent) [Cassell et al., 2000]. This toolkit can generate gestures of different kinds (including iconic gestures) except for metaphoric gestures. Generally, the majority of gesture generation approaches are not considering the effect of emotion (expressed through prosody) on body language, which puts a difficulty towards adapting the generated robot's behavior to human's emotion [Busso and Narayanan, 2007]. In this chapter, we present an extension of our previous work [Aly and Tapus, 2013b] (Chapter 4), which proposes a statistical model for synthesizing adapted head-arm metaphoric gestures to human's prosodic cues. This model has been integrated to the system we are discussing in this chapter, for the purpose of generating an adapted *multimodal* robot's behavior to the emotional content of interaction with the human user.

On the other hand, the correlation between facial expressions and speech had been long recognized in psychological studies [Busso et al., 2004]. The movement of face's muscles and the prosodic cues of speech can change in a synchronized manner in order to communicate different emotions. The single-modal based perception of human's emotion through audio or visual information separately, was discussed in Silva et al. [1997], in which the authors found that some emotions (e.g., sadness and fear) are better characterized by audio information, while other emotions (e.g., happiness and anger) are better characterized by visual information. Chen et al. [1998] discussed the complementarity of both modalities, so that the perception of human's emotion will be ameliorated when both modalities are considered in the same time. These last findings are taken into consideration in the experimental design of our study.

The synthesis and modeling of facial expressions in computer-based applications and 3D

agents received more attention than in human-robot interaction. Parke [1972] developed the first 3D face model that can convey different expressions. Platt and Badler [1981] presented the first model that employs FACS (Facial Action Coding system) in controlling the muscular actions corresponding to facial expressions. Spencer-Smith et al. [2001] developed a more 3D realistic model that can create a stimuli with 16 different FACS action units and determined intensities. Similarly, robots were under continuous study aiming towards allowing them to generate reasonable facial expressions (with less facial flexibility than 3D agents due to mechanical constraints). An early initiative to model facial expressions on robot's face was taken by Breazeal [2003], who developed the robot head Kismet. It uses facial details, like: eyes, mouth, and ears to model facial expressions, such as: anger, happiness, surprise, sadness, and disgust. Breemen et al. [2005] developed the research platform iCat, which can render different facial expressions, such as: sadness, anger, happiness, and fear. Lutkebohle et al. [2010] developed the expressive robot head Flobi that can effectively express emotions, such as: sadness, anger, happiness, fear, and surprise. Beira et al. [2006] developed the complete expressive humanoid robot iCub, which can synthesize a variety of emotions using gestures and facial expressions, including: anger, sadness, surprise, and happiness. In this chapter, we use a highly realistic humanoid robot (ALICE robot) with a special expressive face (Section 5.2.3), for the purpose of generating and evaluating a complete affective multimodal robot's behavior within a human-robot interaction context, which was not sufficiently addressed before in the literature.

The rest of the chapter is structured as following: Section (5.2) presents a detailed illustration for the system architecture, Section (5.3) illustrates the design, the hypotheses, and the scenario of interaction, Section (5.4) provides a description for the experimental results, and finally, Section (5.6) concludes the chapter.

5.2 System architecture

The proposed system is coordinated through different subsystems: (1) Speech dictation system (HTML5 API *multilingual* dictation toolkit), (2) Emotion detection phase, in which some defined keywords are parsed from the dictated speech of human so as to precise an

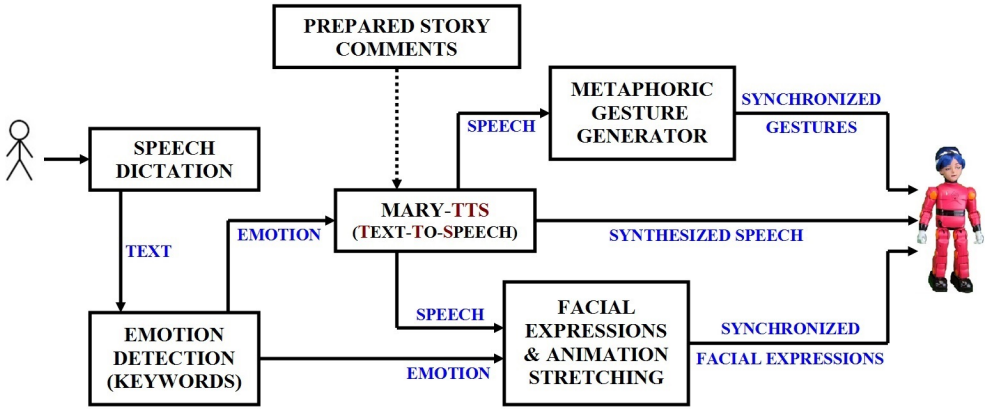


Figure 5-1 – Overview of the emotionally-adapted narrative system architecture

emotional label for the video's content, (3) Mary-TTS engine, which converts the prepared texts (i.e., robot's comments) and the detected emotion in each interaction to a synthesized emotional speech, (4) Metaphoric gesture generator, which maps the synthesized speech to synchronized head-arm metaphoric gestures [Aly and Tapus, 2013b] (Chapter 4), (5) Facial expressions modeling and animation stretching phase, and finally (6) ALICE robot as the test-bed platform in the conducted experiments. An overview of the system architecture is illustrated in Figure (5-1).

5.2.1 Metaphoric Gesture Generator

The generator uses the Coupled (i.e., 2 chains for speech and gestures) Hidden Markov Models (CHMM) [Rabiner, 1989] in order to synthesize head-arm metaphoric gestures, as illustrated in Chapter (4). The motion curves of gesture (i.e., the velocity, acceleration, displacement, and position curves) are segmented by calculating the force, momentum, and kinetic energy of body segments (e.g., the up-arm, low-arm, and hand segments), in addition to calculating the total force of the body. The intersection between these descriptors represents the boundary points of gestures in each body segment. Meanwhile, the pitch-intensity curves of speech are segmented in parallel with gestures in terms of the boundary points of each gesture, the frame time, and the sampling frequency [Aly and Tapus, 2013b]. These segmented patterns of speech and gestures are modeled on the CHMM and are used to train the model, through which new adapted head-arm metaphoric gestures will be generated based on the prosodic cues of a speech-test signal.

5.2.2 Affective Speech Synthesis

The text-to-speech engine (Mary-TTS) is used for the purpose of adding relevant prosodic and accent cues to a preprepared text that summarizes the content of the video under discussion [Schroder and Trouvain, 2003]. This allows the robot to engage in the conversation using - to some extent - adapted emotional speech to the emotional context of the video. The designed vocal patterns are represented using a low-level markup language called MaryXML (which is based on the XML markup language) or using other relatively high-level markup languages, like the SSML (Speech Synthesis Markup Language) [Taylor and Isard, 1997]. The SSML representation offers more vocal design features, like imposing a silence period between words, in addition to an easy control on the characteristics of the pitch contour, baseline pitch, and speech rate (Figure 5-2), which makes it helpful for the emotional vocal patterns' design described in this study. However, the fact that Mary-TTS engine is not prepared yet for efficiently synthesizing emotional speech of different classes in the English language *(to the best of our knowledge, no other vocal engine can)*, makes our proposed vocal design as an *approximate* step towards conveying (even to some extent) the true meaning of the expressed emotion to human. Therefore, the multimodality of the robot's behavior (e.g., when speech and facial expressions are expressed together) is a good solution that could emphasize the general meaning of the expressed behavior, so that each modality enhances the other one.

```
<?xml version="1.0" encoding="UTF-8"?>
- <speak xml:lang="en-US" xsi:schemaLocation="http://www.w3.org/2001/10
  /synthesis http://www.w3.org/TR/speech-synthesis/synthesis.xsd"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="http://www.w3.org/2001/10/synthesis" version="1.0">
   - <p>
      - <prosody rate="-30%" pitch="-4st" contour="(0%,+0st)(100%,-0st)">
           The video's content is so bad
           <break time="0.3s"/>
           innocent people have been attacked by policemen
           <break time="0.3s"/>
           who killed and injured a lot.
        </prosody>
    </p>
</speak>
```

Figure 5-2 – SSML specification of the "sadness" emotion

The designed vocal patterns of the target emotions are summarized in Table (5.1), in which the pitch contours are characterized by sets of parameters inside parentheses (where the first

Emotion	Baseline Pitch	Pitch Contour	Speech Rate	Contour Features			Break Time
				Start	Behavior	End	
Sadness	-4st	(0%,+0st)(100%,-0st)	-30%	Negative	Constant	Negative	Inter/Intra-Sentence
Disgust	+4st	(0%,-5st)(40%,-9st)(75%,-12st)(100%,-12st)	+8%	Negative	Exponential	Negative	Inter-Sentence
Happiness	+2st	(0%,+8st)(30%,+16st)(50%,+14st)(100%,+11st)	+7%	Positive	Parabola	Positive	Inter-Sentence
Anger	+5st	(0%,-18st)(50%,-14st)(75%,-10st)(100%,-14st)	+12%	Negative	Parabola	Negative	Inter-Sentence
Fear	+6st	(0%,+2st)(50%,+5st)(75%,+8st)(100%,+5st)	+7%	Positive	Parabola	Positive	Inter/Intra-Sentence

Table 5.1 – Approximate design of the vocal pattern and the corresponding contour behavior of each target emotion on the standard diatonic scale. Some emotions have used interjections (with tonal stress) in order to emphasize the desired meaning, like: 'Shit' for the "anger" emotion, 'Ugh' and 'Yuck' for the "disgust" emotion, and 'Oh my God' for the "fear" emotion.

parameter in each set followed by *"%"* represents a percentage of the text's duration, while the second parameter represents the corresponding change in the baseline pitch in semitone, which is half of a tone on the standard diatonic scale). The speech rates of the target emotions vary between the rates of the "sadness" emotion (which has the lowest speech rate) and the "anger" emotion (which has the highest speech rate). The inter-sentence break time of each target emotion represents the silence periods between sentences, during which the robot's lips/jaw will show certain expressions that could enhance the expressed emotion (Section 5.2.3). Meanwhile, the intra-sentence break time represents the short silence periods within a sentence, which are necessary for increasing the credibility of the "sadness" and "fear" emotions.

The indicated experimental parameters in Table (5.1) give an example to the prosodic patterns of parts of the texts that Mary-TTS engine should convert to speech in different emotions. The other prosodic patterns of the remaining parts of the texts could differ slightly from the mentioned contour's parameters in order to show some tonal variation through the total of each text.

5.2.3 Face Expressivity

The designed facial expressions corresponding to the prescribed target emotions in this study, are based on the Facial Action Coding System (FACS) [Ekman and Friesen, 1978; Ekman et al., 2002]. Table (5.2) illustrates the FACS coding of each target emotion in our study [Shichuan et al., 2014], in addition to the available equivalent joints in the face of the robot that we used in order to model each expression in the most persuasive manner.

Emotion	FACS Coding	Robot's Face Joint	Additional Body Gestures
Sadness	**Brow Lowerer** + **Lip Corner Depressor** + Inner Brow Raiser + Cheek Raiser + Nasolabial Deepener + Chin Raiser	Left Smile + Right Smile + Brows	*Covering-Eyes Hand* + *Bowing Head* + Narrowing Eyes + Eyes Blinking + Closing Jaw
Disgust	**Nose Wrinkler** + **Upper Lip Raiser** + **Chin Raiser** + Lip Pressor + Brow Lowerer	Jaw + Brows	*Neck Rotation* + *Raising Front-Bent Arms* + Narrowing Eyes
Happiness	**Lip Corner Puller** + **Lips Part** + Cheek Raiser	Left Smile + Right Smile + Jaw	Eyes Blinking
Anger	**Brow Lowerer** + **Lid Tightener** + **Lip Pressor** + Lip Tightener + Upper Lip Raiser + Chin Raiser + Nasolabial Deepener	Jaw + Brows + Eyelids	*Down Head-Shaking* + Short Mouth-Opening
Fear	**Inner Brow Raiser** + **Brow Lowerer** + **Lip Stretcher** + **Lips Part** + Outer Brow Raiser + Upper Lid Raiser + Jaw Drop	*Left Smile* + *Right Smile* + Jaw + Brows + Eyelids	*Mouth-Guard Hand*

Table 5.2 – FACS coding of the target emotions and the corresponding joints in the robot's face, in addition to the other required robot's gestures to emphasize the facial expression's meaning. The bold FACS action units in each emotion represent the observed prototypical units between the subjects in [Shichuan et al., 2014], while the other less common non-bold units are observed with different lower percentages between the subjects. The underlined action units represent the units that have *approximate* corresponding joints in the robot's face.

The complexity behind modeling emotions on the robot's face lies in the absence of the equivalent joints to some FACS descriptors (e.g., cheek raiser, nose wrinkler). Therefore, we imposed experimentally some additional body gestures in order to reduce the negative effect of the missing joints so as to enhance the expressed emotion. These additional gestures do not include -normally- any head gesture (i.e., neck rotation) nor arm-hand gestures, which are being generated by the metaphoric gesture generator explained earlier (Section 5.2.1).[1] However, the combination of the *neck rotation* (i.e., turning the head aside) and the *raising front-bent arms* has been helpful for better expressing the "disgust" emotion (consequently, they got considered as additional supportive gestures for this emotion). This will help give the interacting human - even to some extent - the impression that the robot did not like the context of interaction and considered it disgusting. Similarly, the "fear" emotion, the "anger" emotion, and the "sadness" emotion have been attributed additional *mouth-guard hand* gesture, *down head-shaking*, and *bowing head and covering-eyes hand* gesture respectively, in order to help emphasize their meanings, as indicated in Figure (5-3). On the other hand, the main role of the additional supportive *left smile* and *right smile* robot's

[1]The metaphoric gesture generator (Chapter 4) has the liberty to synthesize the most appropriate gestures based on its own learning algorithm. Therefore, it is *probable* that the previously mentioned supportive head-arm gestures will not be synthesized by the generator during interaction. Consequently, we imposed them at specific moments during speech with a higher priority than the generator's synthesized gestures to make sure of their presence.

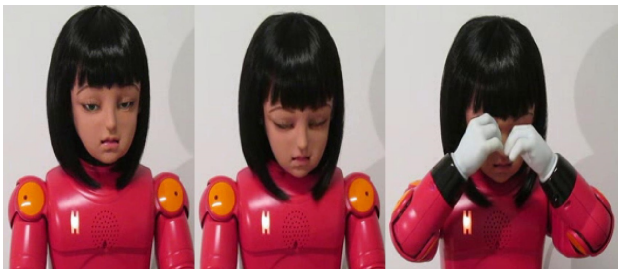
face joints of the “fear" emotion, is to depress a little the corners of the open mouth in order to better reflect the emotion, however they do not have any equivalent FACS descriptor representing the “fear" emotion, as indicated in Table (5.2).

Generally, the modeling of facial expressions on a humanoid robot (even with the expressive ALICE robot) is not an easy task due to the mechanical limitations of the robot’s face (unlike the 3D agents). Therefore, the multimodality of the robot’s behavior is important for interaction, which makes each modality of the combined behavior enhance the other modalities so as to emphasize the conveyed meaning of the expressed emotion to human.

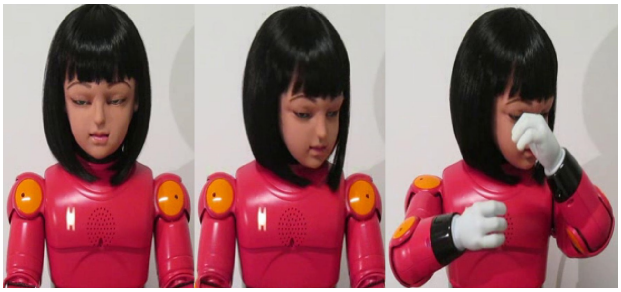
The synchronization between the synthesized emotional speech and the designed facial expressions is controlled by the duration of the generated speech. Figure (5-4) illustrates the XML animation script of the eyelids, in which the 3 control points are characterized in terms of position and time (in milliseconds). In case the duration of the generated speech is longer or shorter than the preliminary duration of the animation, the system calculates easily the new time instant of each control point as a function of the new duration of the generated speech, the preliminary duration of the animation, and the last time instant value of each point. Meanwhile, the position of each control point is kept unchanged.

The segmentation of human’s speech employs the voice activity detection algorithm in order to label and separate between the speech and silence segments. In case the silence period is related to an inter-sentence break time (Section 5.2.2), the robot’s lips/jaw perform certain expressions (e.g., lip corner pulling for the “happiness" emotion) in order to enhance the conveyed meaning of the expressed emotion, as indicated in Figure (5-3). This is due to the mechanical limitations of the robot that do not allow synchronizing the lips with speech while performing an expression with the lips/jaw in the same time. However, in case the silence period is related to an intra-sentence break time (Section 5.2.2), the robot’s jaw is kept opened in the “fear" emotion and closed in the “sadness" emotion during the duration of the short silence period.

On the other hand, the animation of the robot’s lips in a synchronized manner with the segmented speech has encountered a big difficulty when using the 3 servo motors controlling the lips motion (2 motors for the corners and 1 motor for the vertical motion), because they

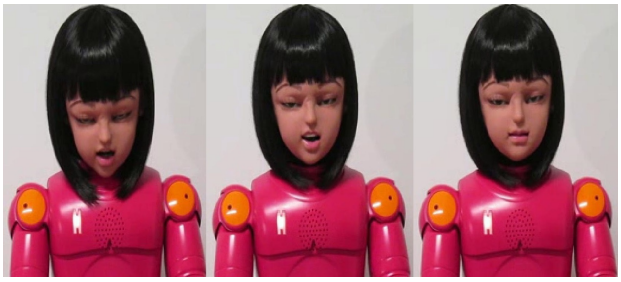


(a) Sadness

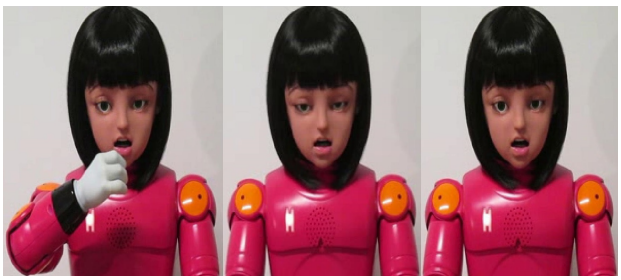


(b) Disgust

(c) Happiness



(d) Anger



(e) Fear

Figure 5-3 – Synthesized facial expressions by ALICE robot

```
<?xml version="1.0"?>
- <Animation>
  - <Version type="Animation">
      <Name>New Animation</Name>
      <Number>1.0</Number>
    </Version>
  - <Channels>
    - <Channel name="Eyelids" id="301">
      - <MotionPaths>
        - <MotionPath name="path">
          - <Version type="Interpolation">
              <Name>C-Spline Interpolation</Name>
              <Number>1.0</Number>
            </Version>
          - <ControlPoints>
            - <ControlPoint>
                <Time>195.0</Time>
                <Position>0.4991511035653651</Position>
              </ControlPoint>
            - <ControlPoint>
                <Time>1270.0</Time>
                <Position>0.9252971137521222</Position>
              </ControlPoint>
            - <ControlPoint>
                <Time>2395.0</Time>
                <Position>0.12224108658743638</Position>
              </ControlPoint>
            </ControlPoints>
          </MotionPath>
        </MotionPaths>
      </Channel>
    </Channels>
</Animation>
```

Figure 5-4 – Eyelids animation script

can not generate a reasonable homogeneous motion when operating together during continuous speech (unlike the 3D conversational agents), in addition to the noise they generate. Alternatively, we used only the motor that controls the vertical motion of the lips. Afterwards, the remote running server of the robot maps the calculated visemes corresponding to the segmented speech to lips motion (where each viseme has a corresponding motion amplitude set experimentally).

5.3 Experimental setup

In this section, we introduce the database used in inducing emotions in human, the experimental hypotheses, the design, and the scenario of interaction between the participant and the humanoid ALICE robot developed by Hanson Robotics (Section 1.2.2).

5.3.1 Database

The database used in this research contains 20 videos inducing the following 6 emotions: sadness, disgust, happiness, anger, fear, and neutral (the “surprise" emotion considered in

Target Emotion	Feature Film
Sadness	**The Champ** - An Officer and a Gentleman
Disgust	**Pink Flamingos** - Maria's Lovers
Happiness	**On Golden Pond** - An Officer and a Gentleman
Anger	**My Bodyguard** - Cry Freedom
Fear	**Halloween** - Silence of the Lambs
Neutral	**Crimes and Misdemeanors** - All the President's Men

Table 5.3 – Target emotions and their corresponding feature films. The main videos used during the experiments were extracted from the bold feature films.

the conducted study in Chapter 4 is not included in the emotion-eliciting database, so that it was excluded and ignored in this chapter). The duration of the videos varies from 29 to 236 seconds, and all of them have been extracted from commercial feature films. The procedures of validating the efficiency of the database in eliciting the target emotions in human, were discussed in Hewig et al. [2005]. During the experiments, we used 12 videos extracted from different films for eliciting the target emotions (which constitute 6 main videos used during the experiments, and 6 standby videos used automatically when the main videos fail to elicit the corresponding target emotions), as indicated in Table (5.3).

5.3.2 Hypotheses

This study aims to test and validate the following hypotheses:

- **H**1: The combination of facial expressions, head-arm metaphoric gestures, and synthesized emotional speech will make the emotional content of interaction more clear to the participant than the interaction conditions that employ less affective cues.

- **H**2: Facial expressions will enhance the expressiveness of the robot's emotion in contrast to the interaction conditions that do not employ facial expressions.

- **H**3: The dynamic characteristics of the robot's head-arm metaphoric gestures will help the participant recognize and distinguish between the target emotions (which is considered as a practical validation for the conducted study in Chapter 4).

5.3.3 Experimental Design

The experimental design of this study is composed of four robot conditions:

- The robot generates a combined multimodal behavior expressed through synchronized head-arm metaphoric gestures, facial expressions, and speech (i.e., condition C1-SFG).
- The robot generates a combined multimodal behavior expressed through synchronized facial expressions and speech (i.e., condition C2-SF).
- The robot generates a combined multimodal behavior expressed through synchronized head-arm metaphoric gestures and speech (i.e., condition C3-SG).
- The robot generates a single-modal behavior expressed only through speech (i.e., condition C4-S).

In order to examine the first hypothesis, the first three conditions were examined (in which facial expressions are accompanied by the additional supportive gestures illustrated in Table 5.2). In this hypothesis, we excluded the conditions of the robot expressing a single-modal behavior through only head-arm metaphoric gestures or facial expressions without speech, in addition to the condition of the robot expressing combined facial expressions and head-arm metaphoric gestures without speech, because they do not match the context of the non-mute human-human interaction. Consequently, the importance of speech in recognizing emotions is measured directly through the questionnaire.

On the other hand, in order to validate the second hypothesis, two conditions were investigated, which are the same as the conditions C2-SF and C4-S. Similarly, in order to validate the third hypothesis, two conditions were tested, which are the same as the conditions C3-SG and C4-S (the condition C2-SF was excluded from validating the third hypothesis, because facial expressions are accompanied by the additional gestures explained earlier in the chapter).

Both of the robot and the interacting human follow a series of short videos that mean to elicit 6 emotions (Section 5.3.1) (Figure 5-5). The different methodologies of emotion

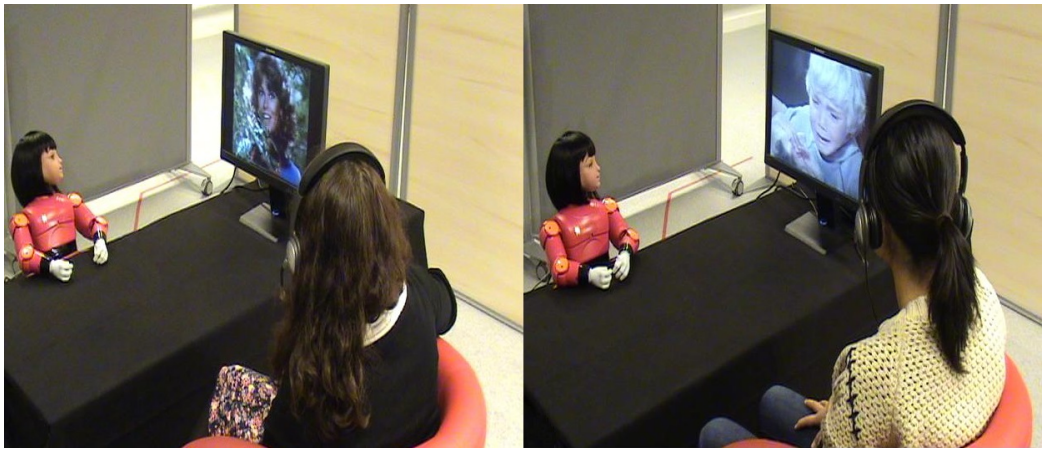


Figure 5-5 – Two participants are interacting with the robot during the “happiness" and “sadness" emotion elicitation experiments

induction and assessment were illustrated in Coan and Allen [2007] and Gil [2009]. The idea behind using videos to elicit certain emotions in human is that they are emotionally convincing, and their role is well indicated in the literature [McHugo et al., 1982; Roberts et al., 2009]. An interesting study about eliciting emotions from films was discussed in Hewig et al. [2005], in which the results proved that the studied target emotions were reasonably recognized. In our study, the scenario of interaction is described as following:

- The robot welcomes the participant and invites him/her to watch some videos so as to have a discussion about.

- The robot asks the participant to express his/her opinion about the content of the video. Afterwards, it parses some expected emotional labels from the dictated comment of the participant, such as: This is *disgusting*!. This helps detect the video’s emotional content so as to trigger an adapted robot’s behavior.

- After listening to the comment of the participant on the video, the robot makes itself a comment accompanied by adapted emotional speech, head-arm metaphoric gestures, and/or facial expressions to the video’s content.

- In case the video elicits a different emotion in the participant from the prescribed target emotion, so that the system parses some keywords that belong mainly to another category of non-target-emotion-referring keywords, the robot will comment through

a *neutral* behavior in order to avoid any emotion-biasing effect. Afterwards, it will invite the participant to watch another video, which means to elicit the same concerned target emotion that was not successfully induced in the participant with the first video.

– The interaction ends for the concerned emotion. Afterwards, the participant starts evaluating the modeled behavior on the robot and the relevance of its emotional content through a Likert questionnaire (in which all questions are presented on a 7-point scale). Whereupon, a new interaction for a new randomly selected emotion starts.

– After the experiments end up, both the robot and the experimenter thank the interacting human for his/her participation.

5.4 Experimental results

The experimental design was based on the between-subjects design, and 60 participant were recruited in order to validate our hypotheses. The participants were uniformly distributed between the four experimental conditions (15 participant (6 female and 9 male) / condition). The recruited participants were ENSTA-ParisTech undergraduate and graduate students and employees whose ages were varying between 20-57 years old ($M = 29.64$, $SD = 9.4$). The background of the participants was non-technical with an average of 33.3%, and technical with an average of 66.7%. 40% of the participants have interacted before with a robot, while 60% of the participants have never interacted with a robot beforehand.

For the first hypothesis, a significant difference was found by ANOVA analysis in the clearness of the robot's emotional behavior expressed through head-arm metaphoric gestures, facial expressions, and speech with respect to the robot's emotional behavior expressed through facial expressions and speech, and the robot's emotional behavior expressed through head-arm metaphoric gestures and speech ($F[2,267] = 9.69$, $p < 0.001$). Tukey's HSD comparisons indicated a significant difference between the robot embodied with head-arm metaphoric gestures, facial expressions, and speech (i.e., condition C1-SFG) from one side, and the robot embodied with facial expressions and speech (i.e., condition

C2-SF) ($p < 0.001$), in addition to the robot embodied with head-arm metaphoric gestures and speech (i.e., condition C3-SG) ($p < 0.001$) from the other side. However, no significant difference was observed between the experimental conditions C2-SF and C3-SG. Moreover, the participants found that the robot's behavior was more expressive in the condition C1-SFG than in the condition C3-SG ($F[1,178] = 13.64$, $p < 0.001$). No significant differences were observed in the participants' ratings regarding the naturalness of the robot's behavior in the conditions C1-SFG, C2-SF, and C3-SG.

For the second hypothesis, the participants found that the robot's behavior expressed though facial expressions and speech (i.e., condition C2-SF) was showing more expressiveness and was more adapted to the content of the interaction than the robot's behavior expressed through speech alone (i.e., condition C4-S) ($F[1,178] = 16.27$, $p < 0.001$). Moreover, the participants considered that facial expressions and speech were synchronized with an average score of $M = 5.9$, $SD = 0.9$. Furthermore, they did not find any significant contradiction between the modalities of the robot's behavior expressed through facial expressions and speech with an average score of $M = 1.8$, $SD = 1.2$. Over and above, they agreed that facial expressions were more expressive than speech with an average score of $M = 4.4$, $SD = 1.5$. Table (5.4) shows that the robot's facial expressions have ameliorated the recognition score of the “anger" emotion in the condition C2-SF with respect to the condition C4-S. This amelioration is related to the encountered difficulties to design a highly persuasive vocal pattern for the “anger" emotion due to the limitations of the Mary-TTS engine (Section 5.2.2). Therefore, the robot's facial expressions have enhanced the affective meaning of speech so as to give the human the feeling that the robot was expressing the “anger" emotion. On the other hand, the robot's facial expressions had a negative influence on the recognition score of the “disgust" emotion in the condition C2-SF with respect to the condition C4-S, which is due to the limited expressivity of the robot's face for this emotion (Section 5.2.3).

For the third hypothesis, the participants considered that the emotional content of the robot's behavior expressed through head-arm metaphoric gestures and speech (i.e., condition C3-SG) was more observable than the emotional content of the robot's behavior

Condition	Emotion					
	Sadness	Disgust	Happiness	Anger	Fear	Neutral
C2-SF	100%	80%	93.3%	92.9%	100%	100%
C3-SG	100%	93.3%	93.3%	92.3%	100%	100%
C4-S	100%	93.3%	93.3%	80%	100%	100%

Table 5.4 – Recognition scores of the target emotions expressed by the robot in 3 different experimental conditions

expressed through speech alone (i.e., condition C4-S) ($F[1,178] = 17.16,\ p = 0.0001$). Furthermore, the participants found that gestures and speech were synchronized with an average score of $M = 6.1$, $SD = 0.7$. At the same time, they agreed that the execution of gestures was fluid with an average score of $M = 5.3$, $SD = 1.01$. Moreover, they considered that gestures were more expressive than speech with an average score of $M = 4.2$, $SD = 1.4$. The emotional content of the robot's head-arm metaphoric gestures was generally recognizable with reasonable scores, as indicated in Table (5.4). However, they have only ameliorated the recognition score of the “anger" emotion in the condition C3-SG with respect to the condition C4-S (similarly to the effect of facial expressions), while the other recognition scores were equal in both conditions. Consequently, the dynamic characteristics of the generated gestures in case of the “anger" emotion, like the high velocity and acceleration, have certainly enhanced the expressive meaning of speech and gave the human the feeling that the robot was angry in a more persuasive manner.

On the other hand, the emotional expressiveness of the robot's behavior was positively perceived in general by the male and female participants in the conditions C2-SF and C3-SG, as indicated in Figures (5-6) and (5-7), respectively. However, the perception of the male participants for the robot's emotional expressiveness in both conditions, was generally higher than the perception of the female participants. The male participants in the condition C2-SF gave higher ratings for the emotions: sadness, disgust, happiness, and fear, meanwhile the female participants gave higher ratings for the emotions: anger and neutral (Figure 5-6). Similarly, the male participants in the condition C3-SG gave higher ratings for the emotions: sadness, disgust, anger, and fear, meanwhile the female participants gave higher ratings for the emotions: happiness and neutral (Figure

5-7). These findings reveal the relatively higher preference of the *male* participants for the emotional expressiveness of the *female* ALICE robot, than the female participants. This gender-based evaluation matches the findings of Siegel et al. [2009], which proved the tendency of the participants to consider the opposite-sex robots as being more credible, engaging, and persuasive. A video showing different synthesized facial expressions, in addition to different interaction experiments with ALICE robot is available at: http://perso.ensta-paristech.fr/~tapus/eng/media.html.

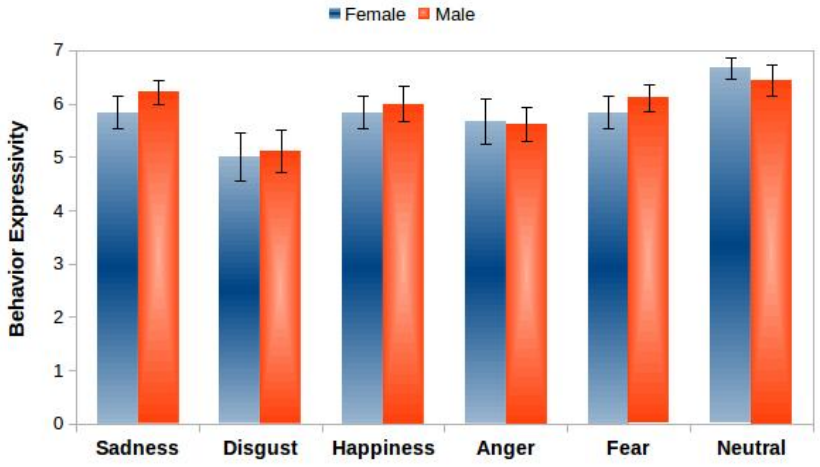


Figure 5-6 – Gender-based evaluation for the emotional expressiveness of the multimodal robot's behavior expressed through combined facial expressions and speech (condition C2-SF). The error bars represent the calculated standard errors.

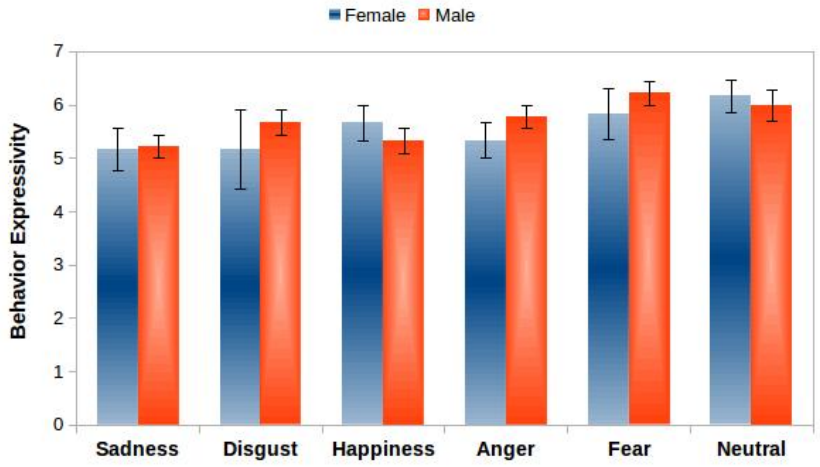


Figure 5-7 – Gender-based evaluation for the emotional expressiveness of the multimodal robot's behavior expressed through combined head-arm metaphoric gestures and speech (condition C3-SG). The error bars represent the calculated standard errors.

5.5 Discussion

In this chapter, we investigated the effect of the generated multimodal robot's behavior on interaction. Moreover, we investigated the effect of facial expressions and head-arm gestures on the expressiveness and clarity of emotion. The proposed context of interaction engages both the robot and the interacting human in limited discussion about the content of a series of videos, which assures a direct human-robot interaction, in addition to a direct evaluation of the human for the generated robot's behavior. This evaluation is based on a Likert questionnaire that poses precise questions about the generated behavior concerning the characteristics of its different modalities of communication, which guarantees that the evaluation of the interacting human for the synthesized multimodal robot's behavior was not inspired by the video under study.

The obtained results validated the important effect of the generated behavior multimodality on enhancing the emotional content of interaction with respect to the generated behaviors that employ less modalities of communication. Besides, the results proved *relatively* the opposite-sex preference principle during human-robot interaction, as indicated in Figures (5-6) and (5-7). However, these findings need a more elaborate study with both a male and a female robots separately, in order to be able to set a global evaluation for this preference, because some other previous studies proved the contrary preference (i.e., the similar-sex preference principle) [Eyssel et al., 2012]. Generally, we believe that both principles are valid, however the tendency of the participants to validate one of them over the other one could be related to both the context of interaction and the task of the robot [Eyssel and Hegel, 2012; Tay et al., 2013], which still needs a further elaborate study.

5.6 Conclusions

This chapter discusses adapting the multimodal robot's behavior to the emotional content of a series of videos eliciting specific target emotions in human within a narrative human-robot interaction. Each interacting human was exposed to one of 4 different ex-

perimental conditions of multimodal/single-modal behaviors during each target emotion elicitation experiment (i.e., each participant was exposed to the same experimental condition in each emotion elicitation experiment). Our proposed system uses Mary-TTS engine in order to generate emotional speech (from the prepared story comments), through which the vocal patterns of the target emotions are designed using the SSML markup language. The metaphoric gesture generator (explained in details in Chapter 4) synthesizes head-arm general gestures based on the prosodic characteristics of speech. On the other hand, the designed and modeled facial expressions on the robot required some additional supportive gestures in order to enhance the conveyed meaning of the expressed emotion.

This chapter validates the role of the robot's behavior multimodality (i.e., combination of facial expressions, gestures, and speech) in increasing the clearness of the emotional content of interaction with respect to the interaction conditions that use less affective cues. Besides, it proves the role of facial expressions in enhancing the expressiveness of the robot's behavior, and the role of the generated gestures (in terms of their dynamic characteristics) in recognizing the target emotions. For the future work, we are interested in increasing the gestural expressivity of the system by integrating additional gesture generators, which can synthesize gestures of other categories (e.g., iconic gestures). Besides, we are interested in ameliorating the emotional content of the synthesized speech so as to make the generated speech more persuasive and natural. This work is under submission in Aly and Tapus [2015a].

Chapter 6

Conclusions

This thesis is about synthesizing an adapted multimodal robot's behavior to human's profile, which is characterized in terms of both the personality and emotion of human. This adapted multimodal robot's behavior sets a basis for enhancing the human-robot long term relationship, in which the robot needs to behave appropriately to the context of interaction. Two humanoid robots (i.e., NAO robot illustrated in Section 1.2.1, and ALICE robot illustrated in Section 1.2.2) have been used for validating the proposed hypotheses in the conducted experimental studies in the thesis.

Chapter (2) focused on developing an online fuzzy-based algorithm for detecting human's emotional state, which considers the evolutionary nature of emotion, so that a group of primary and secondary emotions has been employed in the study. The proposed system was able to successfully precise whether a new detected emotion belongs to one of the learnt clusters so as to get attributed to the corresponding multimodal behavior to the winner cluster, or it constitutes a new cluster that requires synthesizing a new multimodal behavior that matches the context of interaction. This last point is presented in Appendix (C) as a future research direction.

The main encountered problem in creating an online incremental emotion learning system was the fact that people show different amounts of affect in speech according to their personal and cultural characteristics. This could increase the difficulty of making the robot

able to categorize precisely the expressed emotions by pan cultural individuals, which could lead to inappropriate robot's behaviors to the context of interaction. Our proposed system overcame this problem by assuming an uncertain-emotion space within each existing emotion cluster for the data elements whose emotional contents were unconfidently recognized. Once a data element is unconfidently recognized, the robot generates a predefined neutral action in order to avoid any inconsistency in the context of interaction. These uncertain-emotion data elements could, afterwards, create a new cluster, or update an existing emotion cluster upon fulfilling specific criteria.

On the other hand, **Chapter (3)** discussed the importance of personality as a determinant factor for human's verbal and nonverbal behavior. Our study investigated the adaptation of the robot's generated multimodal behavior to the extraversion-introversion personality dimension of the interacting human, which could describe his/her level of sociability (e.g., an extraverted individual tends to be sociable, friendly, fun loving, active, and talkative, while an introverted individual tends to be reserved, inhibited, and quiet). The proposed system in this study employed different subsystems for analyzing human's speech in order to detect his/her extraversion-introversion level, and generating an adapted combined verbal and nonverbal robot's behavior to the detected personality level of the interacting human. The resulting behavior (i.e., personality) adaptation between human and robot could create a more appropriate and attractive interaction, which validates the similarity attraction principle (i.e., individuals are more attracted by others who have similar personality traits) within a human-robot interaction context. Besides, the study proved the important role that gestures play (within an adapted combined speech-gestures robot's behavior) in order to keep engaging the interaction between human and robot with respect to the interaction that employs less affective cues (i.e., interaction through speech only).

The main encountered problems in this system were modeling the dynamic characteristics of the synthesized gestures on the robot so as to reflect appropriately its extraverted and introverted personalities to the interacting human, in addition to keeping the temporal alignment between the generated gestures and speech. These difficulties were taken into account in the system by using motion (and time) control parameters, which adapt

the dynamic characteristics of the generated gestures to the desired personality type and level to show, and maintain the temporal alignment between the synthesized gestures and the corresponding chunks of words (based on the chunks' estimated durations) that will be transformed to speech.

The first part of the thesis ***focused mainly*** on understanding both the emotion and personality of the interacting human so as to make the robot able to behave appropriately. Meanwhile, the following second part of the thesis focused more on developing methodologies for generating an appropriate multimodal robot's behavior (which had been used partially in the conducted study of Chapter 3).

Chapter (4) discussed a methodology for mapping the prosodic characteristics of speech to head-arm metaphoric gestures in order to generate an adapted robot's nonverbal behavior to human's emotion. The proposed system used the Coupled Hidden Markov Models (CHMM) in order to segment prosody patterns and the motion curves of head-arm gestures. The system was trained on an audio-video database covering different emotions so as to allow the CHMM to synthesize a corresponding set of head-arm metaphoric gestures to an observed audio sequence, which helps the robot generate an appropriate body behavior to the interacting human's prosodic characteristics that firmly correlate with his/her emotion.

The main encountered problem in setting up the CHMM model was the adopted conception of gesture segmentation, because people could perceive gesture boundaries in different manners within a continuous motion sequence. Therefore, the traditional approaches of gesture segmentation that use low level descriptors (e.g., velocity and acceleration) could be error-prone, because not all the local minimum points of velocity or acceleration curves represent real gesture boundaries (based on the non-unified manner of the human perception of gesture). This problem indicated the need for a more precise high level representation of gesture. Therefore, we considered the hierarchical construction of the human body and focused on analyzing the dynamic *activity* of the body, which could be measured in terms of the following three high level motion primitives: force, momentum, and kinetic energy of body segments, in addition to the total force of the body. The intersection between these descriptors indicated precisely the boundary points of gestures in each body

segment. Afterwards, the pitch-intensity curves of speech were segmented in parallel with gestures in terms of the calculated boundary points of each gesture. The emotional content of the synthesized gestures was reasonably recognized theoretically. Besides, it was validated practically by human users within human-robot multimodal interaction experiments, as explained later in Chapter (5).

Finally, **Chapter (5)** discussed synthesizing a multimodal adapted robot's behavior to the emotional context of a narrative human-robot interaction (which employs a series of videos in order to elicit the target emotions in the human so as to interact about their storylines). Our study explored and validated the role of the generated behavior multimodality in clarifying the emotional content of interaction, in addition to the positive effect of facial expressions and gestures on the human's perception of interaction. The proposed system in the study used Mary-TTS engine in order to synthesize emotional speech and the expressive face of ALICE robot in order to synthesize emotional facial expressions.

The main encountered problems in this system were synthesizing an emotionally persuasive speech using Mary-TTS engine, which is not prepared yet for efficiently synthesizing emotional speech, in addition to creating credible facial expressions despite the mechanical limitations of the robot's face. For enhancing the synthesized emotional speech, we imposed different interjections with the emotions that were difficult to synthesize by Mary-TTS engine in order to emphasize their desired meanings, like: "anger", "disgust", and "fear" emotions. Besides, we imposed silence periods (i.e., intra-sentence break times) within the synthesized speech in case of the "sadness" and "fear" emotions in order increase their credibility. On the other hand, in order to overcome the lack of some joints in the robot's face that can hinder modeling some facial expressions with credibility, we used additional supportive body gestures that can help reflect - even to some extent - the desired meaning of the emotion.

The ***future research direction*** for this thesis focuses on creating a computational mechanism for synthesizing autonomously a new multimodal robot's behavior, which is slightly presented in Appendix (C). The primary suggested computational model for generating a multimodal action tries to simulate the human cognitive functionalities that understand

both the context and goal of each observed multimodal action in the surrounding environment in order to enhance its incremental learning experience of performing multimodal actions. Consequently, this will make the robot able to synthesize multimodal actions appropriately to the context of interaction whenever required based on the stored cumulative multimodal information in the action memory, which will be - in turn - helpful in building up a successful long term human-robot relationship.

Appendix A

Inverse Kinematics Model of the Arm

After generating the adapted position curves of articulations to human's emotion, the inverse kinematics model of the arm should be formulated in order to calculate the rotation angles of articulations so as to get modeled on the robot.

Considering the transformation matrices describing the position and orientation of the elbow and the end-effector (*relative to the elbow's coordinate frame*) described in Equations (A.1) and (A.2) (where the components of the orientation vectors $\vec{\mathbf{n}}$, $\vec{\mathbf{s}}$, and $\vec{\mathbf{a}}$ are functions of θ_i and α_i, as indicated in Equation 4.8 and Table 4.1) [Asfour and Dillmann, 2003]:

$$T_{Elbow} = \prod_{i=1}^{4} T_i = \begin{pmatrix} n_{x_4} & s_{x_4} & a_{x_4} & p_{x_4} \\ n_{y_4} & s_{y_4} & a_{y_4} & p_{y_4} \\ n_{z_4} & s_{z_4} & a_{z_4} & p_{z_4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{A.1}$$

$$T_{End\,Effector} = \prod_{i=5}^{7} T_i = \begin{pmatrix} n_{x_7} & s_{x_7} & a_{x_7} & p_{x_7} \\ n_{y_7} & s_{y_7} & a_{y_7} & p_{y_7} \\ n_{z_7} & s_{z_7} & a_{z_7} & p_{z_7} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{A.2}$$

Having known the position and orientation of the elbow and the end-effector, it would be

possible to calculate the rotation angles of the left and right arms' articulations directly from the following set of equations (where both $\theta_{Shoulder}$ and θ_{Wrist} have 2 possible solutions):

$$\theta_{Shoulder_{1,2}} = atan2(\pm p_{y_4}, \pm p_{x_4}) \tag{A.3}$$

$$\phi_{Shoulder} = atan2(-p_{z_4}, c(\theta_{Shoulder})\, p_{x_4} + s(\theta_{Shoulder})\, p_{y_4} - L_{Shoulder}) \tag{A.4}$$

$$\begin{aligned}\psi_{Shoulder} =& atan2(-s(\theta_{Shoulder})\, s_{x_4} + c(\theta_{Shoulder})\, s_{y_4}, -s(\phi_{Shoulder})\, c(\theta_{Shoulder})\, s_{x_4} - \\ & s(\phi_{Shoulder})\, s(\theta_{Shoulder})\, s_{y_4} - c(\phi_{Shoulder})\, s_{z_4})\end{aligned} \tag{A.5}$$

$$\begin{aligned}\theta_{Elbow} =& atan2(c(\phi_{Shoulder})\, c(\theta_{Shoulder})\, n_{x_4} + c(\phi_{Shoulder})\, s(\theta_{Shoulder})\, n_{y_4} - \\ & s(\phi_{Shoulder})\, n_{z_4}, c(\phi_{Shoulder})\, c(\theta_{Shoulder})\, a_{x_4} + c(\phi_{Shoulder})\, s(\theta_{Shoulder})\, a_{y_4} - \\ & s(\phi_{Shoulder})\, a_{z_4})\end{aligned} \tag{A.6}$$

$$\theta_{Wrist_{1,2}} = atan2(\pm s_{y_7}, \pm s_{x_7}) \tag{A.7}$$

$$\phi_{Wrist} = atan2(-s_{z_7}, -c(\theta_{Wrist})\, s_{x_7} - s(\theta_{Wrist})\, s_{y_7}) \tag{A.8}$$

$$\psi_{Wrist} = \quad atan2(-s(\theta_{Wrist})\, n_{x_7} + c(\theta_{Wrist})\, n_{y_7}, s(\theta_{Wrist})\, a_{x_7} - c(\theta_{Wrist})\, a_{y_7}) \tag{A.9}$$

Appendix B

Inverse Kinematics Model of the Head

B.1 Forward Kinematics Model of the Head

The human head and neck have both 4 degrees of freedom (DOF): lower-neck pitch, lower-neck roll, upper-neck yaw, and upper-neck pitch rotations. Table (B.1) illustrates the Denavit-Hartenberg parameters of the human head and neck required for the transformation matrix discussed in Equation (4.8) in order to transform the coordinate frame *i-1* to *i* (where the length of the neck is calculated in Section 4.4.2) [Chung, 2009; Milighetti et al., 2011].

$T_{i-1 \rightarrowtail i}$	θ_i	α_i	a_i	d_i
$0 \rightarrowtail 1$	$\theta_{Lower\ Neck} + 90^\circ$	90°	0	0
$1 \rightarrowtail 2$	$\phi_{Lower\ Neck} + 90^\circ$	90°	0	Neck Length
$2 \rightarrowtail 3$	$\theta_{Upper\ Neck} + 90^\circ$	90°	0	0
$3 \rightarrowtail 4$	$\psi_{Upper\ Neck}$	0°	0	0

Table B.1 – Denavit-Hartenberg parameters of the human head and neck

Similarly to the kinematic analysis of the arm discussed in Appendix (A), the inverse kinematics model of the head (which considers also for the neck) should be formulated in order to calculate the 4 rotation angles (having known the orientation of the head), as illustrated in Section (B.2). These calculated rotation angles are used in controlling the motion of the robot's head.

B.2 Inverse Kinematics Model of the Head

Considering the transformation matrix illustrated in Equation (B.1) (where the components of the orientation vectors $\vec{\mathbf{n}}$, $\vec{\mathbf{s}}$, and $\vec{\mathbf{q}}$ are functions of θ_i and α_i, as indicated in Equation 4.8 and Table B.1), the rotation angles of the head could be expressed directly in terms of the orientation vectors, as explained in the following set of equations (where it represents a possible set of inverse kinematics solutions for the 4 joints of the head, however different other solutions could be also applicable):

$$T_{Head} = \prod_{i=1}^{4} T_i = \begin{pmatrix} n_{x_4} & s_{x_4} & q_{x_4} & p_{x_4} \\ n_{y_4} & s_{y_4} & q_{y_4} & p_{y_4} \\ n_{z_4} & s_{z_4} & q_{z_4} & p_{z_4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{B.1}$$

$$\theta_{Lower\ Neck} = \pi \tag{B.2}$$

$$\phi_{Lower\ Neck} = 2 * atan\left(\frac{\sqrt{-q_{x_4}^2 - q_{y_4}^2 + 1} + \sqrt{1 - q_{x_4}^2}}{q_{y_4}}\right) \tag{B.3}$$

$$\theta_{Upper\ Neck} = 2 * atan\left(\frac{\sqrt{1 - q_{x_4}^2} - 1}{q_{x_4}}\right) \tag{B.4}$$

$$\psi_{Upper\ Neck} = acos\left(\frac{-n_{x_4}}{cos(\theta_{Upper\ Neck})}\right) \tag{B.5}$$

Appendix C

Understanding and Generating Multimodal Actions

Physical action understanding in the human brain is considered to be achieved through mirror neurons, which have been discovered first in the premotor and parietal cortices (the F5 and PF areas) of macaque monkeys [Gallese et al., 1996; Fogassi et al., 2005]. Afterwards, different neuroscience studies found evidences that an equivalent mirror neurons system exists in the human brain (in the inferior frontal gyrus, including the Broca's area [Schaffler et al., 1993], which has a major contribution to speech production) [Iacoboni et al., 1999; Ramachandran, 2000; Gazzola and Keysers, 2009]. Mirror neurons get activated when the observer performs a physical action, and when he/she detects others doing the same action. This process requires the observed action to have a goal, so that the observer could estimate the intention of the person performing the action in order to reproduce the same physical action in other similar situations. This discovery had offered a great help towards explaining different high level cognitive phenomena, including understanding physical actions [Rizzolatti and Arbib, 1998; Rizzolatti et al., 2001], and mind reading [Gallese and Goldman, 1998]. Besides, it had led to the "broken mirrors" theory, which revealed some clues that may help researchers develop new approaches to better diagnose autism [Ramachandran and Oberman, 2006]. Moreover, the Wernicke's area located in the superior temporal gyrus of the human brain, is involved in understanding written language through

associating the structure of the written words to their equivalent representations in memory, and similarly with spoken language [Ojemann et al., 1989].

On the other hand, physical action generation is based generally on two learning strategies: *imitation*, in which the observer copies the demonstrator's behavior in order to reach the same result [Whiten and Ham, 1992; Whiten et al., 1996; Whiten, 2002], and *emulation*, in which the observer achieves the same result using his own behavior [Tomasello et al., 1987; Wood, 1989; Tomasello, 1998]. Whiten [2011] distinguished two main subcategories of emulation: (1) end-state result learning (i.e., re-creation of the end of an action sequence by any behavioral means), and (2) affordance learning (i.e., learning about the operating and physical properties of objects through the observation of others when interacting with them, which makes the achievement of similar goals easier without employing imitation). The selection between these two learning strategies (i.e., imitation and emulation) depends mainly on the context. Emulation could be a more convenient strategy than imitation in some contexts due to its flexibility and generality (e.g., when all important causal relationships are clear to the observer - i.e., relationships between causes and effects). Meanwhile, imitation could be more appropriate when these causal relationships are not totally recognized, or when high-fidelity action reproduction is required [Galef, 1992; Heyes, 1993; Tomasello et al., 1993]. On the other hand, speech production associated with the generated physical actions by imitation or emulation, implies intercommunication between different areas in the human brain based on the selected strategy for generating speech, like: repeating a sentence that the observer heard or read, using an existing expression in memory, or formulating a new expression or a group of words based on the accumulated linguistic experience. Repeating a sentence that the observer heard, for example, requires the primary auditory cortex to process the spoken words, then the information travels to the Wernicke's area in order to understand its content, then to the Broca's area in order to formulate the equivalent spoken content of information, and finally to the primary motor cortex, which translates it back into spoken words by controlling the movement of muscles [Ojemann et al., 1989]. Similarly, repeating a sentence read by the observer, requires the primary visual cortex to process the written words, afterwards the processed information travels to the Wernicke's area, then to the Broca's area, and finally to the motor cortex.

On the way for a complete computational cognitive model for understanding multimodal actions, Buchsbaum et al. [2009] discussed an interesting action segmentation approach that segments a sequence of observed body behavior into significant physical actions, through a Bayesian analysis that investigates the inference between causes and effects during action segmentation. Another approach for understanding physical actions was illustrated in Buchsbaum et al. [2011], in which low-level video features were used for the segmentation process. Neural networks have also been employed for action understanding inspired by the human mirror neurons system, and for action generation [Tani, 2003; Tani et al., 2004]. Understanding natural language was- and still is- a challenging topic. It has the objective of extracting all possible information from speech, which necessitates defining the meaning of words and sentences, in addition to precising the corresponding representation of each defined meaning, which makes language understanding as a task-oriented process. Issar and Ward [1993] used a flexible frame-based parser in the development of the CMU's language understanding system. The advantage of this system is that it can deal with the grammatically incorrect formulated sentences, repetitions, etc. Consequently, the system gets able to segment the informative parts of speech in order to directly understand the expressed meaning through semantic analysis. An information retrieval system was discussed in Bennacef et al. [1994], in which a speech recognizer, a semantic analyzer, and a dialog manager were employed. The semantic analyzer performs a case-frame analysis in order to understand the meaning of the processed information. A concept-based approach for understanding language was discussed in Miller et al. [1994] and Levin and Pieraccini [1995], in which language understanding could be considered as a mapping from a sequence of words composing a sentence to a sequence of concepts, where a concept is defined as the smallest meaning-unit.

On the other hand, language generation could be mainly realized whether through predefined language templates that include sentences and words, in addition to some variables that can change the verbosity of the generator's output according to the task, the communicative goal, and human's profile, or through rule-based approaches that use grammatical rules and linguistic constraints in order to calculate the most appropriate verbal output of the system. A common example for the template-based language generation is the weather

forecast generator illustrated in Goldberg et al. [1994]. The main problem associated with the template-based approach is that the generated language is limited linguistically within the prescribed templates of a certain task without big variability [Mcroy et al., 2003; Busemann, 2005], however it is simple to develop. Unlike the template-based approach, the rule-based approach presents a wider linguistic scope for the generated language, so that the linguistic knowledge of the generator could be used for different tasks, even in different languages. However, its relative generality could be a negative point in case the task requires precise information to be given in a certain style [Bateman, 1997; Lavoie and Rambow, 1997]. Similarly, action generation has also faced difficulties in synthesizing a physical behavior relevant to the context of interaction [Gergely, 2003]. Kozima et al. [2002] proposed a human-inspired system for goal emulation, so that it can emulate a goal by its own based on its previous experience. Rudolph et al. [2010] employed a Bayesian network structure in order to store actions as a representation of the resulting effects, which can be used in imitating a physical action, or emulating its goal.

C.1 Cognitive Model Overview

The human cognitive model illustrated in Figure (C-1), is composed of two stages. Stage 1 represents the stage of emotional states detection (Chapter 2), in which an observer decodes and analyzes the contained information in human's speech, reaching an estimation for his/her possible emotional internal state, upon which the observer will generate a corresponding multimodal action. On the other hand, Stage 2 represents an overview of the human cognitive architecture for understanding and generating multimodal actions [Aly and Tapus, 2015b].

Based on the cognitive model discussed in Figure (C-1), an observer understands both the context and goal of each observed multimodal action in the surrounding environment using the mirror neurons and the Wernicke's brain areas. Afterwards, he/she tries to reproduce it by sending the processed information to the synchronization phase for a multimodal temporal alignment, then to the motor cortex that controls the responsible muscles of both speech and gestures generation processes. Whereupon, the aligned multimodal actions

get stored in the action memory. After accumulating enough multimodal interaction experience, and in a moment when an action (i.e., the output of Stage 1) is required to be generated, the action memory will synthesize a multimodal behavior corresponding to the analyzed information in Stage 1, and will send the necessary information to the motor cortex for generating a multimodal action. Besides, the action memory is important during the learning process, because it can offer a base for the emulation process, and same for the Broca's area during speech generation.

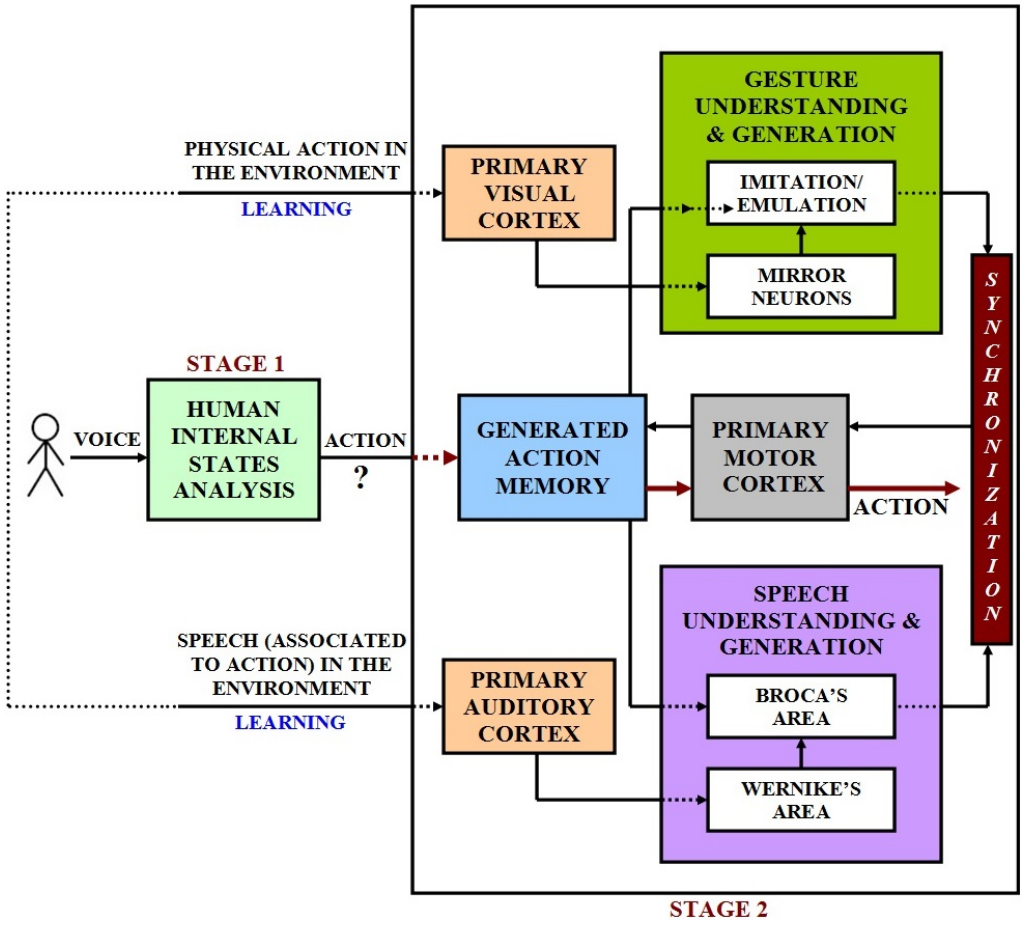


Figure C-1 – Cognitive model for understanding the multimodal actions of humans in the surrounding environment, and for generating multimodal actions corresponding to the detected emotional state

C.2 Computational Model Overview

A preliminary proposed computational model for understanding and generating multimodal actions (i.e., the equivalent computational model to Stage 2 of the human cognitive archi-

tecture discussed above), is illustrated in Figure (C-2). The observed multimodal actions in the environment are captured through appropriate audio and video sensors.

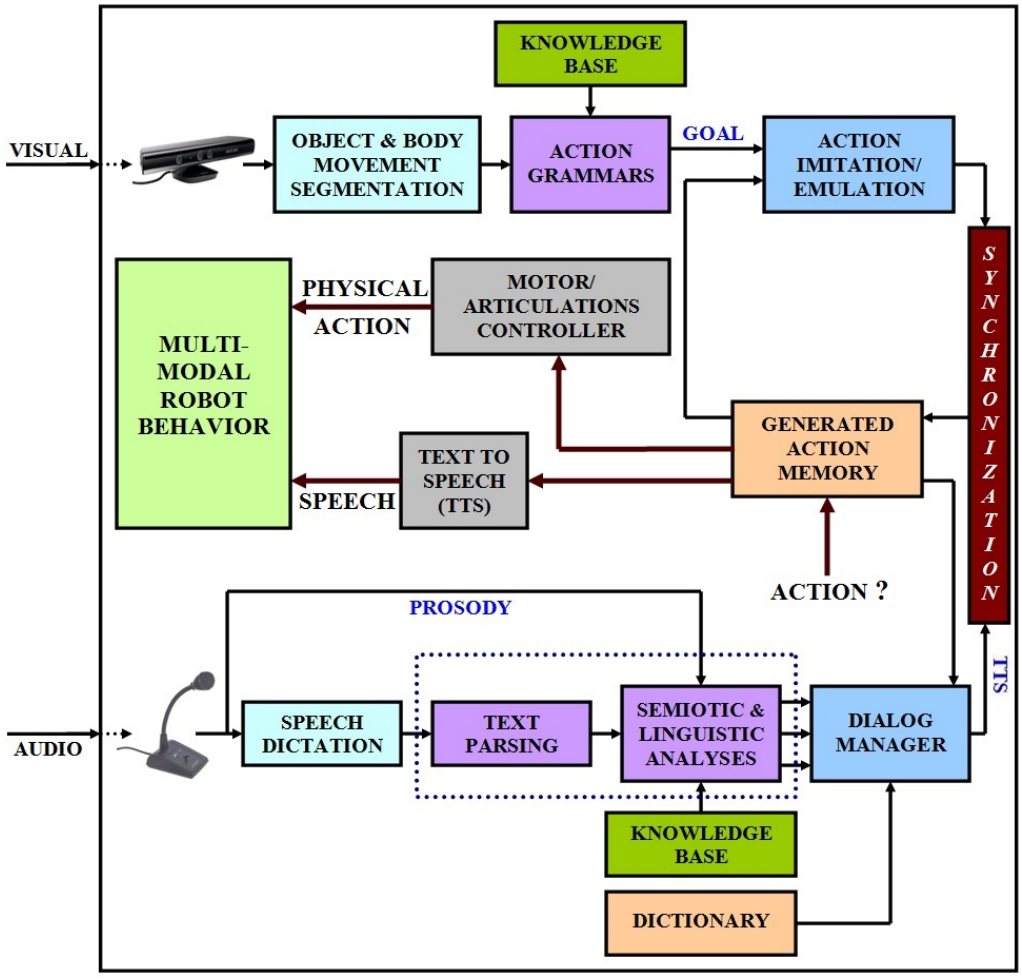


Figure C-2 – Computational model for understanding and generating multimodal actions (Stage 2)

After parsing the text of the dictated speech, semiotic and linguistic analyses are implemented in order to extract the contained pragmatic information, such as speech acts [Searle, 1968, 1969], and the semantic information (i.e., the meanings of words and sentences), and to calculate the interacting human's profile, such as personality traits [Goldberg, 1990; Dang et al., 2012; Aly and Tapus, 2013a]. Afterwards, the dialog manager would generate whether a similar text to the dictated one after understanding its content in the previous step, or a different text expressing the same idea, goal, and context. This process represents the learning phase of the contained information in speech. On the other hand, action

grammars are employed in order to understand the goal of the captured actions [Summers-Stay et al., 2012; Pastra and Aloimonos, 2012]. Therefore, the observed actions could be reproduced by imitation or by emulating its goal. The synchronization phase uses a TTS (text-to-speech) engine in order to calculate the estimated duration of the generated text so as to align it temporally to the generated action. The aligned multimodal actions (learnt by the system) are stored in the action memory, so that the system gets ready for synthesizing a multimodal behavior when an action is required to be generated. This stage is composed of complex subprocesses, and is considered as a future research scope for this thesis.

Bibliography

Aggarwal, J. and Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428–440. 93

Aly, A. and Tapus, A. (2010). Gestures imitation with a mobile robot in the context of human-robot interaction (HRI) for children with autism. In *Proceedings of the 3rd Workshop for Young Researchers on Human-Friendly Robotics (HFR)*, Tubingen, Germany. 22

Aly, A. and Tapus, A. (2011a). Speech-driven arm gestures synthesis in multimodal human-robot interaction. In *Proceedings of the 4th Workshop for Young Researchers on Human-Friendly Robotics (HFR)*, Enschede, The Netherlands. 70

Aly, A. and Tapus, A. (2011b). Speech to head gesture mapping in multimodal human-robot interaction. In *Proceedings of the European Conference on Mobile Robotics (ECMR)*, Orebro, Sweden. 70, 109

Aly, A. and Tapus, A. (2011c). Towards an online voice-based gender and internal state detection model. In *Proceedings of the 6th ACM/IEEE Human-Robot Interaction Conference (HRI)*, Lausanne, Switzerland. 38

Aly, A. and Tapus, A. (2012a). Integrated model for generating non verbal body behavior based on psycholinguistic analysis in human-robot interaction. In *Proceedings of the 1st Post-graduate Conference on Robotics and Development of Cognition (ROBOTDOC-ICANN 2012)*, Lausanne, Switzerland. 58

Aly, A. and Tapus, A. (2012b). An integrated model of speech to arm gestures mapping in human-robot interaction. In *Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, Bucharest, Romania. 70, 109

Aly, A. and Tapus, A. (2012c). Prosody-driven robot arm gestures generation in human-robot interaction. In *Proceedings of the 7th ACM/IEEE Human-Robot Interaction Conference (HRI)*, MA, USA. 70

Aly, A. and Tapus, A. (2012d). Speech to head gesture mapping in multimodal human-robot interaction. In Borangiu, T., Thomas, A., and Trentesaux, D., editors, *Service Orientation in Holonic and Multiagent Manufacturing Control*, volume 402, pages 183–197. Springer-Verlag, Germany. 70, 109

Aly, A. and Tapus, A. (2012e). 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. 52, 102, 112

Aly, A. and Tapus, A. (2013a). 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)*, pages 325–332, Tokyo, Japan. 58, 84, 90, 114, 148

Aly, A. and Tapus, A. (2013b). Prosody-based adaptive metaphoric head and arm gestures synthesis in human robot interaction. In *Proceedings of the 16th IEEE International Conference on Advanced Robotics (ICAR)*, pages 1–8, Montevideo, Uruguay. 63, 70, 71, 72, 109, 111, 114, 116

Aly, A. and Tapus, A. (2014). Towards enhancing human-robot relationship: Customized robot's behavior to human's profile. In *Proceedings of the AAAI Fall Symposium on AI for Human-Robot Interaction (AI-HRI)*, VA, USA. 15, 23

Aly, A. and Tapus, A. (2015a). Multimodal adapted robot's behavior synthesis within a narrative human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany. 131

Aly, A. and Tapus, A. (2015b). An online fuzzy-based approach for human emotions detection: An overview on the human cognitive model of understanding and generating multimodal actions. In Mohammed, S., Moreno, J., Kong, K., and Amirat, Y., editors, *Intelligent Assistive Robots: Recent Advances in Assistive Robotics for Everyday Activities*, volume 106. Springer-Verlag, Switzerland. 52, 112, 146

Aly, A. and Tapus, A. (2015c). Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human-robot interaction. *Autonomous Robots*. 84

Andre, E., Rist, T., Mulken, S., Klesen, M., and Baldes, S. (2000). The automated design of believable dialogues for animated presentation teams. In *S. Prevost, J. Cassell, J. S. and Churchill, E., editors,Embodied conversational agents*, pages 220–255, MA, USA. MIT Press. 56

Ang, M. and Tourassis, V. (1987). Singularities of Euler and roll-pitch-yaw representations. *IEEE Transactions on Aerospace and Electronic Systems*, 23(3):317–324. 95

Angelov, P. (2002). *Evolving rule-based models: A tool for design of flexible adaptive systems*. Heidelberg, Germany. 42, 44, 45

Asfour, T. and Dillmann, R. (2003). 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. 97, 139

Bachorowski, J. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57. 112

Badler, N., Costa, M., Zhao, L., and Chi, D. (2000). To gesture or not to gesture: What is the question? In *Proceedings of the International Conference on Computer Graphics*, Geneva, Switzerland. 99

Bailenson, J. and Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science*, 16:814–819. 60

Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Personality and Social Psychology*, 70:614–636. 36, 38

Bargh, J., Chen, M., and Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Personality and Social Psychology*, 71:230–244. 60

Barrick, M. and Mount, M. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Pshychology*, 44:1–26. 61, 62

Bateman, A. (1997). Enabling technology for multilingual natural language generation: The KMPL development. *Natural Language Engineering*, 3:15–55. 146

Beattie, G. and Sale, L. (2012). Do metaphoric gestures influence how a message is perceived? The effects of metaphoric gesture-speech matches and mismatches on semantic communication and social judgment. *Semiotica*, 2012(192):77–98. 81

Beira, R., Lopes, M., Praga, M., Santos-Victor, J., Bernardino, A., Metta, G., Becchi, F., and Saltaren, R. (2006). Design of the robot-cub (iCub) head. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 94–100, USA. 115

Bennacef, S., Bonnea-Maynard, H., Gauvain, J., Lamel, L., and Minker, W. (1994). A spoken language system for information retrieval. In *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan. 145

Bennewitz, M., Faber, F., Joho, D., and Behnke, S. (2007). FRITZ – a humanoid communication robot. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Jeju, South Korea. 90

Bernieri, F. (1988). Coordinated movement and rapport in teacher-student interactions. *Nonverbal Behavior*, 12:120–138. 60

Bevacqua, E., Mancini, M., and Pelachaud, C. (2004). Speaking with emotions. In *AISB Convention: Motion, Emotion and Cognition*, Leeds, UK. University of Leeds. 57

Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, NY, USA. 32

Bobick, A. and Wilson, A. (1995). A state based technique for the summarization and recognition of gesture. In *Proceedings of the 5th International Conference on Computer Vision*, MA, USA. 100

Brand, M. (1999). Voice puppetry. In *Proceedings of the ACM SIGGRAPH Asia*, pages 21–28, NY, USA. 91

Breazeal, C. (2003). Towards sociable robots. *Robotics and Autonomous Systems*, 42:167–175. 115

Breazeal, C. and Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12:83–104. 30, 36

Breemen, A., Yan, X., and Meerbeek, B. (2005). iCat: An animated user-interface robot with personality. In *Proceedings of the 4th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, Utrecht, The Netherlands. 115

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In *Proceedings of the ACM SIGGRAPH Asia*, pages 353–360, NY, USA. 91

Buchsbaum, D., Canini, K., and Griffiths, T. (2011). Segmenting and recognizing human action using low-level video. In *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. 145

Buchsbaum, D., Griffiths, T., Gopnik, A., and Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, The Netherlands. 145

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech (http://database.syntheticspeech.de). In *Proceedings of Interspeech*, Lisbon, Portugal. 35, 38

Busemann, S. (2005). Ten years after: An update on TG/2 (and friends). In *Proceedings of the 10th European Natural Language Generation Workshop*, Aberdeen, Scotland. 146

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech, and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, pages 205–211, NY, USA. 114

Busso, C. and Narayanan, S. (2007). Interrelation between speech and facial gestures in emotional utterances: A single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2331–2347. 114

Byrne, D. and Griffit, W. (1969). Similarity and awareness of similarity of personality characteristic determinants of attraction. *Experimental Research in Personality*, 3:179–186. 59

Cahn, J. (1990a). Generating expression in synthesized speech. Master's thesis, MIT Media Lab., MA, USA. 30

Cahn, J. (1990b). The generation of affect in synthesized speech. *The American Voice I/O Society*, 8:1–19. 112

Campbell, L. and Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *Proceedings of the 5th International Conference on Computer Vision*, MA, USA. 100

Cappella, J. and Planalp, S. (1981). Talk and silence sequence in informal conversations III: Interspeaker influence. *Human Communication Research*, 7(2):117–132. 60

Casasanto, D. and Lozano, S. (2007). The meaning of metaphorical gestures. In Cienki, A. and Muller, C., editors, *Metaphor and Gesture*. John Benjamins, Amsterdam, The Netherlands. 89

Cassell, J. and Bickmore, T. (2003). Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. In *User Modeling and User-Adapted Interaction*, volume 13, pages 89–132. 56

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., and Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E., editors, *Embodied Conversational Agents*, pages 29–63. MIT Press, MA, USA. 57, 90, 114

Cassell, J., Vilhjálmsson, H., and Bickmore, T. (2001). BEAT: The behavior expression animation toolkit. In *Proceedings of the SIGGRAPH*, pages 477–486. 15, 63, 67, 89, 90, 114

Chartrand, T. and Bargh, J. (1999). The chameleon effect: The perception-behavior link and social interaction. *Personality and Social Psychology*, 76(6):893–910. 60

Chen, L., Huang, T., Miyasato, T., and Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, pages 366–371, Nara, Japan. 114

Chiu, S. (1994). Fuzzy model identification based on cluster estimation. *Intelligent and Fuzzy Systems*, 2(3):267–278. 33, 40, 41

Chklovski, T. and Pantel, P. (2004). VERBOCEAN: Mining the web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain. 66

Chung, H. (2009). *Optimization-based dynamic prediction of 3D human running*. PhD thesis, University of Iowa, IA, USA. 141

Cienki, A. (2000). Gesture, metaphor, and thinking for speaking. In *Proceedings of the 5th Conference on Conceptual Structure, Discourse, and Language (CSDL)*, CA, USA. 89

Clavel, C., Devillers, L., Plessier, J., Ach, L., Morel, B., and Martin, J. (2012). Combinaisons d'expressions vocales, faciales et posturales des émotions chez un agent animé: Perception par les utilisateurs. *Technique et Science Informatiques*, 31:533–564. 31

Coan, J. and Allen, J. (2007). *The handbook of emotion elicitation and assessment (Series in affective science)*. Oxford University Press, NY, USA. 125

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33:497–505. 65

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297. 35

Courgeon, M., Martin, J., and Jacquemin, C. (2008). MARC: A multimodal affective and reactive character. In *Proceedings of the AFFINE Workshop on Affective Interaction in Natural Environments, in parallel with the 10th International Conference on Multimodal Interfaces (ICMI)*, Chania, Crete. 57

Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32. 112

Cristianini, N. and Shawe-Taylor, J. (2000). Introduction to support vector machines. *Cambridge University Press*. 38

Dang, H., Aly, A., and Tapus, A. (2012). Robot personality design for an appropriate response to the human partner. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*, Paris, France. 148

DeCarolis, B., Pelachaud, C., Poggi, I., and Steedman, M. (2004). APML, a mark-up language for believable behavior generation. In Prendinger, H. and Ishizuka, M., editors, *Life-Like Characters: Tools, Affective Functions and Applications*, pages 65–85. Springer-Verlag, Germany. 90

Deng, Z., Busso, C., Narayanan, S., and Neumann, U. (2004). Audio-based head motion synthesis for avatar-based telepresence systems. In *Proceedings of the ACM SIGMM Workshop on Effective Telepresence (ETP)*, pages 24–30. 91

Dewaele, J. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, 49(3):509–544. 63

Dicaprio, N. (1983). *Personality theories: A guide to human nature*. Holt, Rinehart and Wilson, NY, USA. 57

Dijkstra, P. and Barelds, D. (2008). Do people know what they want: A similar or complementary partner? *Evolutionary Psychology*, 6(4):595–602. 59

Dunn, J. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Cybernetics*, 3:32–57. 32

Edgington, M. (1997). Investigating the limitations of concatenative synthesis. In *Proceedings of Eurospeech*, Greece. 113

Ekman, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. Pergamon Press, NY, USA. 33

Ekman, P. (1979). About brows: Emotional and conversational signal. In Cranach, M. V., Foppa, K., Lepenies, W., and Ploog, D., editors, *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, UK. 112

Ekman, P. and Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98. 88, 113

Ekman, P. and Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, CA, USA. 118

Ekman, P., Friesen, W., and Ellsworth, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? In Ekman, P., editor, *Emotion in the Human Face*, pages 39–55. Cambridge University Press, NY, USA. 33, 93

Ekman, P., Friesen, W., and Hager, J. (2002). *Facial action coding system [E-book]*. Research Nexus, UT, USA. 118

Eriksson, J., Matarić, M., and Winstein, C. (2005). Hands-off assistive robotics for post-stroke arm rehabilitation. In *Proceedings of the IEEE International Conference on Rehabilitation Robotics (ICORR)*, pages 21–24, IL, USA. 58

Eyereisen, F. and Lannoy, J. (1991). *Gestures and speech: Psychological investigations*. Cambridge University Press, Cambridge, UK. 88, 112

Eysenck, H. (1953). *The structure of human personality*. London, UK. 57, 81

Eysenck, H. (1991). Dimensions of personality: 16, 5 or 3? Criteria for a taxonomic paradigm. *Personality and Individual Differences*, 12:773–790. 57, 81

Eysenck, H. and Eysenck, S. (1968). *Manual: Eysenck personality inventory*. Educational and Industrial Testing Service, CA, USA. 81

Eyssel, F. and Hegel, F. (2012). (s)he's got the look: Gender-stereotyping of social robots. *Applied Social Psychology*, 42(3):2213–2230. 130

Eyssel, F., Kuchenbrandt, D., Hegel, F., and de Ruiter, L. (2012). Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *Proceedings of the 21th IEEE International Symposium in Robot and Human Interactive Communication (RO-MAN)*, pages 851–857, Paris, France. 130

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press, MA, USA. 66

Finin, T., Joshi, A., and Webber, B. (1986). Natural language interactions with artificial experts. In *Proceedings of the IEEE,10(2)*, pages 921–938. 56

Fogassi, L., Ferrari, P., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organization to intention understanding. *Science*, 308:662–667. 143

Fontaine, J., Scherer, K., Roesch, E., and Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12):1050–1057. 51

Forbes-Riley, K. and Litman, D. (2007). Investigating human tutor responses to student uncertainty for adaptive system development. In *Lecture Notes in Computer Science*, volume 4738, pages 678–689. 56

Forbes-Riley, K., Litman, D., and Rotaru, M. (2008). Responding to student uncertainty during computer tutoring: An experimental evaluation. In *Lecture Notes in Computer Science*, volume 5091, pages 60–69. 56

Furnham, A. (1990). *Language and personality*. In Giles, H. and Robinson, W. (Eds.), Handbook of Language and Social Psychology. Winley. 63

Galef, B. (1992). The question of animal culture. *Human Nature*, 3:157–178. 144

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119:593–609. 143

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind reading. *Trends in Cognitive Sciences*, 2:493–500. 143

Gath, I. and Geva, A. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:773–781. 32

Gazzola, V. and Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex*, 19:1239–1255. 143

Gergely, G. (2003). What should a robot learn from an infant? Mechanisms of action interpretation and observational learning in infancy. *Connection Science*, 15:191–209. 146

Gil, S. (2009). Comment étudier les émotions en laboratoire? *Revue Electronique de Psychologie Sociale*, 4:15–24. 125

Giles, H. and Powesland, P. (1978). Speech style and social evaluation. In Erickson, F., editor, *Language in Society*, pages 428–433. Cambridge University Press, Cambridge, UK. 60

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368. 65

Goddard, N. (1994). Incremental model based discrimination of articulated movement from motion sequences. In *Proceedings of the IEEE Computer Society Workshop on Motion of Non Rigid and Articulated Objects*. 100

Goldberg, E., Driedger, N., and Kittredge, R. (1994). Using natural language processing to produce weather forecasts. *IEEE Intelligent Systems and their Applications*, 9:45–53. 146

Goldberg, L. (1990). An alternative description of personality: The Big-Five factor structure. *Personality and Social Psychology*, 59:1216–1229. 57, 64, 148

Goldberg, L. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, 7:7–28. 57, 64

Grandjean, D., Sander, D., and Scherer, K. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition*, 17(2):484–495. 51

Grosz, B. (1983). TEAM: A transportable natural language interface system. In *Proceedings of the Conference on Applied Natural Language Processing*, pages 39–45, CA, USA. 56

Gueguen, N. (2007). *100 petites experiences en psychologie de la seduction*. Dunod, Paris, France. 61

Gump, B. and Kulik, J. (1997). Stress, affiliation, and emotional contagion. *Personality and Social Psychology*, 72:305–319. 61

Gustafsson, D. and Kessel, W. (1979). Fuzzy clustering with a fuzzy covariance matrix. In *Proceedings of the IEEE Conference on Decision and Control (CDC) including the 17th Symposium on Adaptive Processes*, pages 761–766, CA, USA. 32

Hall, E. (1966). *The hidden dimension*. Doubleday, NY, USA. 62

Hartmann, B., Mancini, M., and Pelachaud, C. (2002). Formational parameters and adaptive prototype instantiation for MPEG-4 compliant gesture synthesis. In *Proceedings of the Computer Animations*, Geneva, Switzerland. IEEE Computer Society Press. 57

Hassin, R. and Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Personality and Social Psychology*, 78:837–852. 63

Hewig, J., Hagemann, D., Seifert, J., Gollwitzer, M., Naumann, E., and Bartussek, D. (2005). A revised film set for the induction of basic emotions. *Cognition and Emotion*, 19(7):1095–1109. 123, 125

Heyes, C. (1993). Imitation, culture, and cognition. *Animal Behavior*, 46:999–1010. 144

Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, 286:2526–2528. 143

Iida, A. and Campbell, N. (2003). Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders. *Speech Technology*, 6(4):379–392. 113

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics. *Human-Computer Studies*, 53:251–267. 59, 63

Issar, S. and Ward, W. (1993). CMU's robust spoken language understanding system. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin, Germany. 145

Izard, C. (1971). *Face of emotion*. Appleton, NY, USA. 33, 34

Izard, C. (1977). *Human emotions*. Plenum Press, NY, USA. 34

J-Campbell, L., Gleason, K., Adams, R., and Malcolm, K. (2003). Interpersonal conflict, agreeableness, and personality development. *Personality*, 71(6):1059–1085. 62

Jensen, C., Farnham, S., Drucker, S., and Kollock, P. (2000). The effect of communication modality on cooperation in online environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 470–477, NY, USA. 90

Jones, C. and Deeming, A. (2008). Affective human-robot interaction. In Peter, C. and Beale, R., editors, *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, pages 175–185. Springer-Verlag, Germany. 31

Jung, C., Hull, R., and Baynes, H. (1976). *Psychological types*. Princeton University Press, NJ, USA. 79

Kahol, K., Tripathi, P., and Panchanathan, S. (2003). Gesture segmentation in complex motion sequences. In *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain. 99, 100

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica*, 32:100–125. 88

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Key, M., editor, *The Relationship of Verbal and Nonverbal Communication*, pages 207–227. Mouton Publishers, The Hague, The Netherlands. 70, 88, 113

Kendon, A. (1982). The study of gesture: Some observations on its history. *Recherches Semiotique/Semiotic Inquiry*, 2:45–62. 88

Kendon, A. (1983). The study of gesture: Some remarks on its history. In Deely, J., editor, *Semiotics 1981*, pages 153–164. Springer-Verlag, NY, USA. 113

Kendon, A. (1994). Does gesture communicate? A review. *Research on Language and Social Interaction*, 2(3):175–200. 88

Kipp, M., Neff, M., Kipp, K., 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*, Paris, France. Springer-Verlag. 90

Kopp, S., Bergmann, K., and Wachsmuth, I. (2008). Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production. *Semantic Computing*, 2(1):115–136. 57, 90, 114

Kopp, S., Krenn, B., Marsella, S., Marshall, A., Pelachaud, C., Pirker, H., Thórisson, K., and Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The behavior markup language. *Intelligent Virtual Agents*, pages 205–217. 56

Kopp, S. and Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52. 90, 114

Kozima, H., Nakagawa, C., and Yano, H. (2002). Emergence of imitation mediated by objects. In *Proceedings of the 2nd International Workshop on Epigenetic Robotics*, Edinburgh, Scotland. 146

Kroemer, K., Kroemer, H., and Kroemer-Elbert, K. (1994). *Ergonomics: How to design for ease and efficiency*. Prentice Hall. 96

Lafrance, M. (1982). Posture mirroring and rapport. In Davis, M., editor, *Interaction Rhythms: Periodicity in Commutative Behavior*, pages 279–298. Human Sciences Press, NY, USA. 60

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, IL, USA. 89

Lakoff, G. and Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to western thought*. University of Chicago Press, IL, USA. 89

Lavoie, B. and Rambow, O. (1997). A fast and portable realizer for text generation. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP)*. 66, 146

Le, Q., Huang, J., and Pelachaud, C. (2012). 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. 90, 114

Le, Q. and Pelachaud, C. (2012). Generating co-speech gestures for the humanoid robot NAO through BML. In *Proceedings of the 9th international conference on Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, pages 228–237. 56

Leary, T. (1957). *Interpersonal diagnosis of personality*. Ronald Press, NY, USA. 59

Lee, C. and Xu, Y. (1996). Online interactive learning of gestures for human-robot interfaces. In *Proceedings of the IEEE International Conference on Robotics and Automation*, MN, USA. 100

Lee, K., Peng, W., Jin, S.-A., and Yan, C. (2006). Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human robot interaction. *Communication*, 56:754–772. 59

Leuwerink, K. (2012). A robot with personality: Interacting with a group of humans. In *Proceedings of the 16th Twente Student Conference on IT*, Enschede, The Netherlands. 62

Leva, P. (1996). Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters. *Biomechanics*, 29(9):1223–1230. 96

Levin, E. and Pieraccini, R. (1995). Concept-based spontaneous speech understanding system. In *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, GA, USA. 145

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*, CA, USA. 91

Levine, S., Theobalt, C., and Koltun, V. (2009). Real-time prosody-driven synthesis of body language. In *Proceedings of the ACM SIGGRAPH Asia*, NY, USA. 91, 93

Lippa, R. and Dietz, J. (2000). The relation of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Nonverbal Behavior*, 24:25–43. 62

Luo, P., Ng-Thow-Hing, V., and Neff, M. (2013). An examination of whether people prefer agents whose gestures mimic their own. In *Intelligent Virtual Agents: Lecture Notes in Computer Science*, volume 8108, pages 229–238. 61

Lutkebohle, I., Hegel, F., Schulz, S., Hackel, M., Wrede, B., Wachsmuth, S., and Sagerer, G. (2010). The Bielefeld anthropomorphic robot head Flobi. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3384–3391, AK, USA. 115

Mairesse, F. and Walker, M. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37. 15, 63, 66

Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Artificial Intelligence Research (JAIR)*, pages 457–500. 63, 64

Mamdani, E. and Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. *Man-Machine Studies*, 7(1):1–13. 31

Mancini, M. and Pelachaud, C. (2008). Distinctiveness in multimodal behaviors. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 159–166. 57

Maurer, R. and Tindall, J. (1983). Effects of postural congruence on client's perception of counselor empathy. *Counseling Psychology*, 30:158–163. 60

McHugo, G., Smith, C., and Lanzetta, J. (1982). The structure of self-reports of emotional responses to film segments. *Motivation and Emotion*, 6:365–385. 125

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press, IL, USA. 56, 81, 88, 89, 113

McNeill, D. (2000). *Language and gesture*. Cambridge University Press, Cambridge, UK. 56, 81, 89, 113

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press, IL, USA. 81

Mcroy, S., Channarukul, S., and Ali, S. (2003). An augmented template based approach to text realization. *Natural Language Engineering*, 9:381–420. 146

Mehl, M., Gosling, S., and Pennebaker, J. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Personality and Social Psychology*, 90:862–877. 64

Mey, J. (2001). *Pragmatics: An introduction*. Blackwell Publishers. 88

Milighetti, G., Vallone, L., and De-Luca, A. (2011). Adaptive predictive gaze control of a redundant humanoid robot head. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3192–3198, CA, USA. 141

Miller, S., Bobrow, R., Schwartz, R., and Ingria, R. (1994). Statistical language processing using hidden understanding models. In *Proceedings of the Human Language Technology Workshop*, NJ, USA. 145

Montero, J., Gutierrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S., and Pardo, J. (1998). Emotional speech synthesis: from speech database to TTS. In *Proceedings of the International Conference on Spoken Language Processing*, pages 923–925, Sydney, Australia. 36, 38

Moon, Y. and Nass, C. (1996). How real are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research*, 23:651–674. 63

Morris, L. (1979). *Extraversion and introversion: An interactional perspective*. Hemisphere Publishing Corporation, NY, USA. 57

Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Acoustical Society of America*, 93(2):1097–1108. 30

Murray, I. and Arnott, J. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, 16(4):369–390. 112

Murray, J. (1990). Review of research on the myers-briggs type indicator. *Perceptual and Motor Skills*, 70:1187–1202. 57

Myers-Briggs, I. and Myers, P. (1980). *Gifts differing: Understanding personality type*. Davies-Black Publishing, CA, USA. 57

Nakajima, H., Morishima, Y., Yamada, R., Brave, S., Maldonado, H., Nass, C., and Kawaji, S. (2004). Social intelligence in a human-machine collaboration system: Social responses to agents with mind model and personality. *Japanese Society for Artificial Intelligence*, 19(3):184–196. 62

Nakajima, H., Nass, S., Yamada, R., Morishima, Y., and Kawaji, S. (2003). The functionality of human-machine collaboration systems-mind model and social behavior. In *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, pages 2381–2387, VA, USA. 62

Nass, C. and Lee, M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency attraction. *Experimental Psychology: Applied*, 7:171–181. 58, 59, 63

Nean, A., 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, pages 2013–2016, FL, USA. 104

Neff, M., Kipp, M., Albrecht, I., and Seidel, H. (2008). Gesture modeling and animation based on a probabilistic recreation of speaker style. *ACM Transactions on Graphics*, 27(1):1–24. 90

Ng-Thow-Hing, V., Luo, P., and Okita, S. (2010). Synchronized gesture and speech production for humanoid robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan. 56, 90

Niewiadomski, R., Hyniewska, S., and Pelachaud, C. (2009). Modeling emotional expressions as sequences of behaviors. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, pages 316–322, Amsterdam, The Netherlands. 57

Ojemann, G., Ojemann, J., Lettich, E., and Berger, M. (1989). Cortical language localization in left, dominant hemisphere: An electrical stimulation mapping investigation in 117 patients. *Neurosurgery*, 71:316–326. 144

Park, E., Jin, D., and Del-Pobil, A. (2012). The law of attraction in human-robot interaction. *Advanced Robotic Systems*, 9(35). 59

Parke, F. (1972). Computer generated animation of faces. In *Proceedings of the ACM Annual Conference*, volume 1, pages 451–457, NY, USA. 115

Pastra, K. and Aloimonos, Y. (2012). The minimalist grammar of action. *Philosophical Transactions of the Royal Society Biological Sciences*, 367:103–117. 149

Peca, A., Tapus, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., and David, D. (2012). Exploratory study: Children's with autism awareness of being imitated by Nao robot. In *Proceedings of the 1st International Conference on Innovative Technologies for Autism Spectrum Disorders. ASD: Tools, Trends and Testimonials (ITASD)*, Valencia, Spain. 22

Pelachaud, C. (2005). Multimodal expressive embodied conversational agents. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pages 683–689, NY, USA. 90, 113

Pell, M. and Kotz, S. (2011). On the time course of vocal emotion recognition. *PLoS ONE*, 6(11). 112

Pennebaker, J. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Personality and Social Psychology*, 77:1296–1312. 64

Peter, C. and Beale, R. (2008). *Affect and emotion in human-computer interaction: From theory to applications*. Springer Verlag, Germany. 31

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: Features and algorithms. *Human-Computer Studies*, 59(1):157–183. 31

Pittam, J. (1994). *Voice in social interaction: An interdisciplinary approach*. Sage, CA, USA. 63

Plagenhoef, S., Evans, F., and Abdelnour, T. (1983). Anatomical data for analyzing human motion. *Research Quarterly for Exercise and Sport*, 54:169–178. 96, 97

Platt, J. (1998). Fast training of support vector machines using sequential mininal optimization. Technical report, Microsoft Research MSR-TR-98-14. 38

Platt, S. and Badler, N. (1981). Animating facial expressions. *Computer Graphics*, 15:245–252. 115

Plutchik, R. (1991). *The emotions*. University Press of America, MD, USA. 15, 33, 34, 35, 93, 112

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286. 92, 104, 106, 116

Rabiner, L., Atal, B., and Sambur, M. (1977). LPC prediction error: Analysis of its variation with the position of the analysis frame. *IEEE Transactions on Systems Man, and Cybernetics*, 25:434–442. 37

Ramachandran, V. (2000). Mirror neurons and imitation learning as the driving force behind "the great leap forward" in human evolution. *Edge*, 69. 143

Ramachandran, V. and Oberman, L. (2006). Broken mirrors: A theory of autism. *Scientific American*, 295:62–69. 143

Reeves, B. and Nass, C. (1996). The media equation. *University of Chicago Press*. 59

Reiter, E. and Dale, R. (2000). *Building natural language generation systems*. Cambridge University Press, Cambridge, UK. 65

Rezek, I. and Roberts, S. (2000). 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. 92, 104, 105

Rezek, I., Sykacek, P., and Roberts, S. (2000). Coupled hidden Markov models for biosignal interaction modeling. In *Proceedings of the 1st International Conference on Advances in Medical Signal and Information Processing (MEDSIP)*, pages 54–59, UK. 92, 104, 105, 106

Riggio, R. and Friedman, H. (1986). Impression formation: The role of expressive behavior. *Personality and Social Psychology*, 50:421–427. 63

Rizzolatti, G. and Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences*, 21:188–194. 143

Rizzolatti, G., Fogassi, L., and Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, 2:661–670. 143

Roberts, D., Narayanan, H., and Isbell, C. (2009). Learning to influence emotional responses for interactive storytelling. In *Proceedings of the AAAI Spring Symposium on Intelligent Narrative Technologies*, pages 95–102, Stanford University, USA. 125

Rong, J., Li, G., and Chen, Y.-P. (2008). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Management*, 45:315–328. 38

Roy, D. and Pentland, A. (1996). Automatic spoken affect analysis and classification. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, VT, USA. 30

Rudolph, M., Muhlig, M., Gienger, M., and Bohme, H. (2010). Learning the consequences of actions: Representing effects as feature changes. In *Proceedings of the International Symposium on Learning and Adaptive Behavior in Robotic System*. 146

Russell, J. (1980). A circumplex model of affect. *Personality and Social Psychology*, 39(6):1161–1178. 51

Sargin, M., Yemez, Y., Erzin, E., and Tekalp, A. (2008). Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1330–1345. 91

Sauter, D., Eisner, F., Calder, A., and Scott, S. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, 63(11):2251–2272. 112

Schaffler, L., Luders, H., Dinner, D., Lesser, R., and Chelune, G. (1993). Comprehension deficits elicited by electrical stimulation of Broca's area. *Brain*, 116:695–715. 143

Scherer, K. (1979). Language and personality. In Scherer, K. and Giles, H., editors, *Social Markers in Speech*, pages 147–209. Cambridge University Press, Cambridge, UK. 63

Scherer, K. (2000). Psychological models of emotion. In Borod, J., editor, *The Neuropsychology of Emotion*, pages 137–162. Oxford University Press, NY, USA. 51

Scherer, K. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40:227–256. 112

Scherer, K., Schorr, A., and Johnstone, T. (2001). *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, NY, USA. 51

Schroder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system Mary: A tool for research, development, and teaching. *Speech Technology*, 6(4):365–377. 113, 117

Searle, J. (1968). Austin on locutionary and illocutionary acts. *The Philosophical Review*, 77:405–424. 148

Searle, J. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, Cambridge, UK. 148

Selfhout, M., Burk, W., Branje, S., Denissen, J., Aken, M., and Meeus, W. (2010). Emerging late adolescent friendship networks and Big Five personality traits: A social network approach. *Personality*, 78(2):509–538. 62

Shichuan, D., Yong, T., and Martinez, A. (2014). Compound facial expressions of emotion. In *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, volume 111, pages 1454–1462. 19, 118, 119

Shroder, M. (2009). Expressive speech synthesis: Past, present, and possible futures. In Tao, J. and Tan, T., editors, *Affective Information Processing*, pages 111–126. Springer-Verlag, UK. 88, 112

Siegel, M., Breazeal, C., and Norton, M. (2009). Persuasive robotics: The influence of robot gender on human behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2563–2568, MO, USA. 129

Silva, L. D., Miyasato, T., and Nakatsu, R. (1997). Facial emotion recognition using multimodal information. In *Proceedings of IEEE International Conference on Information, Communications, and Signal Processing (ICICS)*, volume 1, pages 397–401, Singapore. 114

Slaney, M. and McRoberts, G. (1998). Baby ears: A recognition system for affective vocalizations. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Seattle, USA. 30

Sondhi, M. (1968). New methods of pitch extraction. *IEEE Transactions Audio and Electroacoustics*, 16:262–266. 37

Spencer-Smith, J., Wild, H., Innes-Ker, A., Townsend, J., Duffy, C., Edwards, C., Ervin, K., Merritt, N., and Paik, J. (2001). Making faces: Creating three-dimensional parameterized models of facial expression. *Behavior Research Methods, Instruments, and Computers*, 33(2):115–123. 115

Stent, A., Prasad, R., and Walker, M. (2004). Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 79–86, NJ, USA. 66

Sugeno, M. (1985). *Industrial applications of fuzzy control*. Elsevier Science Publishing Corporation, NY, USA. 31

Sullivan, H. (1953). *The interpersonal theory of psychiatry*. Norton, NY, USA. 59

Summers-Stay, D., Teo, C., Yang, Y., Fermuller, C., and Aloimonos, Y. (2012). Using a minimal action grammar for activity understanding in the real world. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, MD, USA. 149

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systemsm Man, and Cybernetics*, 15:116–132. 31, 33

Talkin, D. (1995). A robust algorithm for pitch tracking. In Kleijn, W. and Paliwal, K., editors, *Speech Coding and Synthesis*, pages 497–518. Elsevier. 36, 37

Tani, J. (2003). Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. *Neural Networks*, 16:11–23. 145

Tani, J., Ito, M., and Sugita, Y. (2004). Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks*, 17:1273–1289. 145

Tapus, A. and Aly, A. (2011). User adaptable robot behavior. In *Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS)*, PA, USA. 70

Tapus, A. and Matarić, M. (2008). Socially assistive robots: The link between personality, empathy, physiological signals, and task performance. In *Proceedings of the AAAI Spring Symposium on Emotion,Personality and Social Behavior*, CA, USA. 57, 59

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., and David, D. (2012a). Children with autism social engagement in interaction with Nao, an imitative robot-a series of single case experiments. *Interaction Studies*, 13(3):315–347. 22

Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., and David, D. (2012b). Children with autism social engagement in interaction with Nao, an imitative robot-a series of single case experiments. In *Proceedings of the 2nd International Conference on Innovative Research in Autism (IRIA)*, Tours, France. 22

Tapus, A., Tapus, C., and Matarić, J. (2008). User-robot personality matching and robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics, Special Issue on Multidisciplinary Collaboration for Socially Assistive Robotics*, 1(2):169–183. 62

Tay, B., Park, T., Jung, Y., Tan, Y., and Wong, A. (2013). When stereotypes meet robots: The effect of gender stereotypes on people's acceptance of a security robot. In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics: Understanding Human Cognition*, volume 8019, pages 261–270. Springer-Verlag, Germany. 130

Taylor, P. and Isard, A. (1997). SSML: A speech synthesis markup language. *Speech Communication*, 21:123–133. 117

Tomasello, M. (1998). Emulation learning and cultural learning. *Behavior and Brain Science*, 21:703–704. 144

Tomasello, M., Davis-Dasilva, M., Camak, L., and Bard, K. (1987). Observational learning of tool use by young chimpanzees and enculturated chimpanzees. *Human Evolution*, 2:175–183. 144

Tomasello, M., Savage-Rumbaugh, E., and Kruger, A. (1993). Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child Development*, 64:1688–1705. 144

Tomkins, S. (1984). Affect theory. In Scherer, K. and Ekman, P., editors, *Approaches to Emotion*, pages 163–195. Hillsdale, NJ:Erlbaum. 34

Van-Baaren, R., Holland, R., Steenaert, B., and Van-Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Experimental Social Psychology*, 39:393–398. 60

Vapnik, V. (1998). *Statistical learning theory*. Wiley-Blackwell. 32

Vinacke, W., Shannon, K., Palazzo, V., and et al., L. B. (1988). Similarity and complementarity in intimate couples. *Genetic, Social, and General Psychology Monographs*, 114(1):51–76. 60

Voeffra, C. (2011). Emotion-sensitive human-computer interaction: State of the art. Seminar on Emotion Recognition. 31

Vogel, K. and Vogel, S. (1986). L'interlangue et la personnalite de l'apprenant. *Applied Linguistics*, 24(1):48–68. 65

Vogt, T. and Andre, E. (2006). Improving automatic emotion recognition from speech via gender differentiation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*. 31

Vroomen, J., Collier, R., and Mozziconacci, S. (1993). Duration and intonation in emotional speech. In *Proceedings of Eurospeech*, Germany. 113

Wahlster, W. and Kobsa, A. (1989). User models in dialog systems. pages 4–34, Germany. Springer Verlag. 56

Wang, T., Shum, H., Xu, Y., and Zheng, N. (2001). Unsupervised analysis of human gestures. In *Proceedings of the IEEE Pacific Rim Conference on Multimedia*, Beijing, China. 100

Webb, J. (1972). Interview synchrony: An investigation of two speech rate measures. In Siegman, A. and Pope, B., editors, *Studies in Dyadic Communication*, pages 115–133. Pergamon Press, NY, USA. 60

Whiten, A. (2002). Imitation of sequential and hierarchical structure in action: Experimental studies with children and chimpanzees. In Dautenhahn, K. and Nehaniv, C., editors, *Imitation in Animals and Artifacts*, pages 191–209. MIT Press, MA, USA. 144

Whiten, A. (2011). The scope of culture in chimpanzees, humans, and ancestral apes. *Philosophical Transactions of the Royal Society*, 366:935–1187. 144

Whiten, A., Custance, D., Gomez, J., Teixidor, P., and Bard, K. (1996). Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). *Comparative Psychology*, 110:3–14. 144

Whiten, A. and Ham, R. (1992). On the nature and evolution of imitation in the animal kingdom: Reappraisal of a century of research. *Advances in the Study of Behavior*, 21:239–283. 144

Williams, J. (1971). Personal space and its relation to extraversion-introversion. *Canadian Journal of Behavioural Science*, 3(2):156–160. 62

Windhouwer, D. (2012). The effects of the task context on the perceived personality of a Nao robot. In *Proceedings of the 16th Twente Student Conference on IT*, Enschede, The Netherlands. 61

Winter, D. (2009). *Biomechanics and motor control of human movement*. John Wiley and Sons Ltd. 96

Wood, D. (1989). Social interaction as tutoring. In Bornstein, M. and Bruner, J., editors, *Interaction in Human Development*, pages 59–80. Psychology Press, NJ, USA. 144

Woods, S., Dautenhahn, K., Kaouri, C., Boekhorst, R., and Koay, K. (2005). Is this robot like me? Links between human and robot personality traits. In *Proceedings of the 5th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Tsukuba, Japan. 57

Woods, S., Dautenhahn, K., Kaouri, C., Boekhorst, R., Koay, K., and Walters, M. (2007). Are robots like people? Relationships between participant and robot personality traits in human-robot interaction studies. *Interaction Studies*, 8(2):281–305. 58

Yager, R. and Filev, D. (1992). Approximate clustering via the mountain method. Technical report, MII 1305, Machine Intelligence Institute, Iona College, New Rochelle, NY, USA. 32

Yager, R. and Filev, D. (1993). Learning of fuzzy rules by mountain clustering. In *Proceedings of the SPIE Conference on Applications of Fuzzy Logic Technology*, pages 246–254, Boston, USA. 32

Young, K. (1927). *Source book for social psychology*. A.A. Knopf, NY, USA. 79

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338–353. 31

Zadeh, L. (1973). Outline of a new approach to the analysis of complex systems and decision processes. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(1):28–44. 31

Zatsiorsky, V. and Seluyanov, V. (1979). Mass inertial characteristics of human body segments and their relationship with anthropometric landmarks (in russian). *Voprosy Antropologii*, 62:91–103. 96

Zukerman, I. and Litman, D. (2001). Natural language processing and user modeling: Synergies and limitations. In *User Modeling and User-Adapted Interaction*, volume 11, pages 129–158. 56

Curriculum Vitae

Amir ALY was born in Egypt in 1983. He obtained his first B.Sc. degree in control and measurement engineering from Benha University, Egypt, in 2006. In 2007, he received a postgraduate scholarship from the French Government. Afterwards, he obtained a second B.Sc. degree in automation and signal processing from ENSISA, France, in 2009. In 2010, he obtained his M.Sc. degree in image and sound processing from Pierre and Marie Curie University (UPMC-Paris 6), France.

He started his work in ENSTA-ParisTech as a Ph.D. researcher in 2010. His research focuses on understanding and generating multimodal behaviors in Human-Robot Interaction (HRI) contexts. He successfully defended his thesis on December 16, 2014, where the dissertation was titled: *Towards an Interactive Human-Robot Relationship: Developing a Customized Robot's Behavior to Human's Profile*. The jury members appreciated the quality of his work and the wide scope of the covered topics during his Ph.D., therefore they decided to bestow on him the highest degree of honours. Last but not least, he was a subject of inclusion in the highly reputable encyclopedias: the "Marquis Who's Who in the World" and the "2000 outstanding intellectuals of the 21th century" of Cambridge university, in 2014.

Publications

- **Journal Papers**

- Aly, A., Tapus, A., "Towards an Intelligent System for Generating an Adapted Verbal and Nonverbal Combined Behavior in Human-Robot Interaction", Autonomous Robots (under review), 2015.

- Tapus, A., Peca, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., and David, D., "Children with Autism Social Engagement in Interaction with Nao, an Imitative Robot - A Series of Single Case Experiments", Interaction Studies Journal, 13(3), pp 315-347, 2012.

– **Conference Papers**

- Aly, A., and Tapus, A., "Multimodal Adapted Robot's Behavior Synthesis within a Narrative Human-Robot Interaction", in Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (under review), 2015.

- Aly, A., and Tapus, A., "Prosody-Based Adaptive Metaphoric Head and Arm Gestures Synthesis in Human-Robot Interaction", in Proceedings of the 16th IEEE International Conference on Advanced Robotics (ICAR), Montevideo, Uruguay, 2013.

- Aly, A., and Tapus, A., "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 Human-Robot Interaction Conference (HRI), Tokyo, Japan, 2013.

- Aly, A., and Tapus, A., "Towards an Online Fuzzy Modeling for Human Internal States Detection", in Proceedings of the 12th IEEE International Conference on Robotics and Automation (ICARCV), Guangzhou, China, 2012.

- Peca, A., Tapus, A., Aly, A., Pop, C., Jisa, L., Pintea, S., Rusu, A., and David, D., "Exploratory Study: Children's with Autism Awareness of Being Imitated by Nao Robot", in Proceedings of the 1st International Conference on Innovative Technologies for Autism Spectrum Disorders. ASD: Tools, Trends and Testimonials (ITASD), Valencia, Spain, 2012.

- Aly, A., and Tapus, A., "An Integrated Model of Speech to Arm Gestures Mapping in Human-Robot Interaction", in Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM12), Bucharest, Romania, 2012 ***(BEST PAPER AWARD)***.

- Aly, A., and Tapus, A., "Speech to Head Gesture Mapping in Multimodal Human-Robot Interaction", in Proceedings of the European Conference on Mobile Robotics (ECMR), Orebro, Sweden, 2011.

– **Book Chapters**

– Aly, A., and Tapus, A., “An Online Fuzzy-Based Approach for Human Emotions Detection: An Overview on the Human Cognitive Model of Understanding and Generating Multimodal Actions“, In Intelligent Assistive Robots, Series: Springer Tracts on Advanced Robotics (STAR), S. Mohammed et al.(eds.), vol. 106, Switzerland, 2015.

– Aly, A., and Tapus, A., ”Speech to Head Gesture Mapping in Multimodal Human-Robot Interaction”, In Service Orientation in Holonic and Multiagent Manufacturing Control, Series: Springer Tracts in Studies in Computational Intelligence, T. Borangiu et al.(eds.), vol. 402, pp. 183-197, 2012.

– **Short Conference and WS Papers**

– Aly, A., and Tapus, A., “Towards Enhancing Human-Robot Relationship: Customized Robot’s Behavior to Human’s Profile”, in Proceedings of the AAAI Fall Symposium on AI for Human-Robot Interaction, Virginia, USA, 2014.

– Aly, A., and Tapus, A., “How to Infer the Intention of Human while Interacting with a Robot?”, in Proceedings of the 5th Workshop for Young Researchers on Human- Friendly robotics (HFR), Brussels, Belgium, 2012.

– Aly, A., and Tapus, A., “Integrated Model for Generating Non Verbal Body Behavior Based on Psycholinguistic Analysis in Human-Robot Interaction”, in Proceedings of the 1st Post-graduate Conference on Robotics and Development of Cognition (ROBOTDOC-ICANN 2012), Lausanne, Switzerland, 2012.

– Dang, T.H.H, Aly, A., and Tapus, A., “Robot Personality Design for an Appropriate Response to the Human Partner“, in Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Paris, France, 2012.

– Tapus, A., Peca, A., Aly, A., and al., ”Children with Autism Social Engagement in Interaction with Nao, an Imitative Robot - A Series of Single Case

Experiments“, in Proceedings of the 2nd International Conference on Innovative Research in Autism (IRIA), Tours, France, 2012.

– Aly, A., and Tapus, A., ”Prosody-Driven Robot Arm Gestures Generation in Human Robot Interaction“, in Proceedings of the 7th ACM/IEEE Human-Robot Interaction Conference (HRI), Boston, USA, 2012.

– Aly, A., and Tapus, A., ”Speech-Driven Arm Gestures Synthesis in Multimodal Human-Robot Interaction“, in Proceedings of the 4th Workshop for Young Researchers on Human-Friendly Robotics (HFR), Enschede, Netherlands, 2011.

– Tapus, A., and Aly, A., ”User Adaptable Robot Behavior“, in Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS), Pennsylvania, USA, 2011.

– Aly, A., and Tapus, A., ”Towards an Online Voice-Based Gender and Internal State Detection Model“, in Proceedings of the 6th ACM/IEEE Human-Robot Interaction Conference (HRI), Lausanne, Switzerland, 2011.

– Aly, A., and Tapus, A., ”Gestures Imitation with a Mobile Robot in the Context of Human Robot Interaction for Children with Autism“, in Proceedings of the 3rd Workshop for Young Researchers on Human-Friendly Robotics, Tubingen, Germany, 2010.