An Online Fuzzy-Based Approach for Human Emotions Detection: An Overview on the Human Cognitive Model of Understanding and Generating Multimodal Actions

Amir Aly and Adriana Tapus

ENSTA ParisTech, Robotics and Computer Vision Lab,
828 Boulevard des Maréchaux, 91120 Palaiseau, France
{amir.aly,adriana.tapus}@ensta-paristech.fr
http://cogrob.ensta-paristech.fr

Abstract. An intelligent robot needs to be able to understand human emotions, and to understand and generate actions through cognitive systems that operate in a similar way to human cognition. In this chapter, we mainly focus on developing an online incremental learning system of emotions using Takagi-Sugeno (TS) fuzzy model. Additionally, we present a general overview for understanding and generating multimodal actions from the cognitive point of view. The main objective of this system is to detect whether the observed emotion needs a new corresponding multimodal action to be generated in case it constitutes a new emotion cluster not learnt before, or it can be attributed to one of the existing actions in memory in case it belongs to an existing cluster.

Keywords: Plutchik model, Takagi-Sugeno (TS) fuzzy model, Potential calculation, Cluster centers.

1 Introduction

The fast developing human-robot interaction (HRI) applications require the robot to be capable of dealing appropriately with different and varying situations as human does. This objective necessitates the robot to have high level cognitive functions that can make it able to detect the emotional state of the interacting human in order to generate a corresponding action. The robot's ability to generate an appropriate action requires an understanding of the context, of the environment, and of human intentions and performed actions combined with its own accumulated experience.

Physical action understanding in human's 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 [1,2]. Afterwards, different neuroscience studies found evidences that an equivalent mirror neurons system exists in human's brain (in the inferior frontal gyrus, including the Broca's area [3], which has a major contribution to speech production) [4,5,6].

S. Mohammed et al. (eds.), *Intelligent Assistive Robots*,
Springer Tracts in Advanced Robotics 106, DOI: 10.1007/978-3-319-12922-8_7

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 [7,8], and mind reading [9]. Besides, it had led to the "broken mirrors" theory, which revealed some clues that may help researchers develop new approaches to better diagnose autism [10]. Moreover, the Wernicke's area located in the superior temporal gyrus of human's 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 [11].

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 [12,13,14], and *emulation*, in which the observer achieves the same result using his own behavior [15,16,17]. Whiten in [18] 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 (e.g., 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 [19,20,21]. On the other hand, speech production associated with the generated physical actions by imitation or emulation, implies intercommunication between different areas in human's 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 [11]. 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, Buchsbau et al., in [22] discussed an action segmentation ap-

proach 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 [23], 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 [24,25]. 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 in [26] 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 [27], 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 [28,29], 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 the 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 [30]. 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 [31,32], 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 [33,34]. Similarly, action generation has also faced difficulties in synthesizing a physical behavior relevant to the context of interaction [35]. Kozima et al., in [36] 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., in [37] 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.

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 affective states from voice signal. The significance of prosody in conveying emotions is illustrated in [38,39]. The authors discussed a comparative study about the variation of some relevant parameters (such as pitch and voice quality) in case of different emotional states. Moreover, Cahn in [39] 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 in [40] illustrated a spoken affect analysis system that can detect speaker's approval versus speaker's disapproval in child-directed speech. Similarly, Slaney and McRoberts in [41] proposed a system that can recognize prohibition, praise, and attentional bids in infant-directed speech. Breazeal and Aryananda in [42] 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 emotion recognition from voice through gender differentiation was described in [43]. 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 system can recognize human emotions in order to generate an adapted behavior so as to maintain maximum engagement with human [44]. Similarly, emotion-based applications appear in different areas with robots, like: entertainment, education, and general services [45,46].

On the other hand, the importance of using fuzzy logic in modeling complex systems has been increased gradually in the last decade. It imitates human logic by using a descriptive and imprecise language in order to cope with input data. Zadeh in [47,48] 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. This validates the necessity of modeling emotional states 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 [49] and Sugeno [50]. Mamdani in [49] 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 in [51,50] 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 [52,53]. 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 [54], then improved by Bezdek [52], 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 in [55] developed the classical Fuzzy C-means algorithm by 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 in [56] described an unsupervised extension of the algorithm illustrated in [55] (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 [57,58] 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 [59] 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 chose to use the subtractive clustering algorithm in order to identify the parameters of the TS fuzzy model [59,50].

The rest of the chapter is structured as following: Section 2 presents an overview for the cognitive system, Section 3 illustrates a general overview for the basic and complex emotions, Section 4 discusses the offline detection of human's emotional states, Sections 5 and 6 overview the subtractive clustering and Takagi-Sugeno fuzzy model, Section 7 describes the online updating of Takagi-Sugeno fuzzy model, Section 8 provides a description of the results, and finally Section 9 concludes the chapter.

2 Cognitive System Overview

The human cognitive model illustrated in Figure (1), is composed of two stages. Stage 1 represents the stage of emotional states detection, in which an observer decodes and analyzes the contained information in human speech, reaching to an estimation for his possible emotional emotional state, upon which the observer will generate a corresponding multimodal action. In this chapter, we will focus on this stage, and we will try to develop a computational model with the same functionality based on fuzzy logic. On the other hand, Stage 2 represents an overview of human cognitive architecture for understanding and generating multimodal actions. An observer learns the context and goal of each observed multimodal action in the surrounding environment. He/she tries to reproduce it by sending the processed information to the synchronization phase for multimodal temporal alignment, then to the motor cortex that controls the responsible muscles of both speech and gesture generation processes. Thereupon, 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 multimodal action generation. Similarly, the action memory is important during the learning process, so that it can offer a base for the emulation process, and same for the Broca's area during speech generation.

A preliminary proposed computational model for understanding and generating multimodal actions (i.e., the equivalent computational model to Stage 2 of the human cognitive architecture discussed above), is illustrated in Figure (2). The observed multimodal actions in the environment are captured through appropriate audio and video channels. After parsing the text of the dictated speech, semiotic and linguistic analyses are implemented in order to extract the contained pragmatic information (i.e., speech acts [60,61]) and the semantic

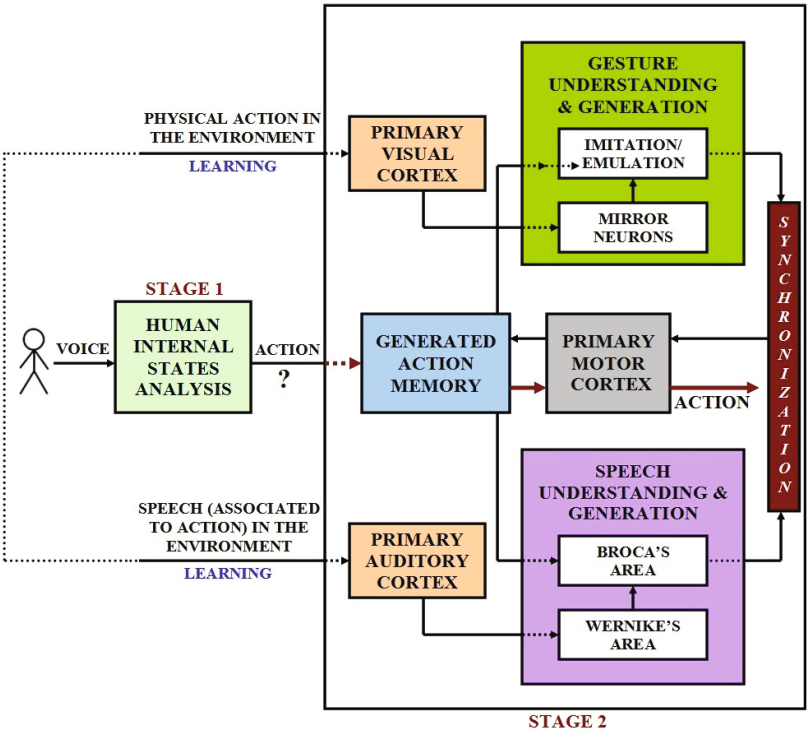

Fig. 1. Cognitive model for understanding multimodal actions of people in the surrounding environment, and generating multimodal actions corresponding to detected emotional states.

information (i.e., the meanings of words and sentences), and to calculate the interacting human's profile (i.e., personality traits [62,63]). Afterwards, the dialog manager can generate a similar text to the dictated one after understanding its content in the previous step, or it can generate a different text but 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 [64,65]. Thereupon, the observed action could be learnt 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 multi-complex sub-processes, and is considered as a future direction for the current research focusing on Stage 1.

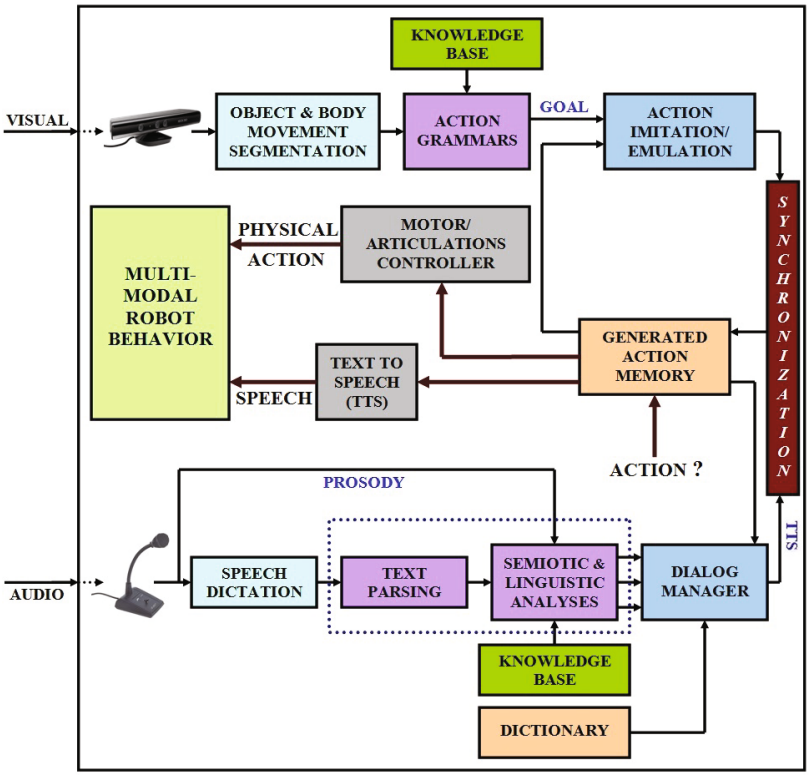

Fig. 2. Computational model for understanding and generating multimodal actions (Stage 2)

3 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 [66,67]. 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 [67].

Different emotion theories identified relatively small sets of fundamental or basic emotions, which are meant to be fixed and universal to all humans (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 in [68,69] 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 resolve the problem discussed in the research of Izard [66], 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 in [70] 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 in [71] 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 [67]. 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 (3). The 8 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 (different polarity; i.e., 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 this research.

4 Offline Detection of Emotional States

In this research, we investigated the performance of the offline classification system using the Support Vector Machine (SVM) algorithm [72], 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 contained 5 complex emotions (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

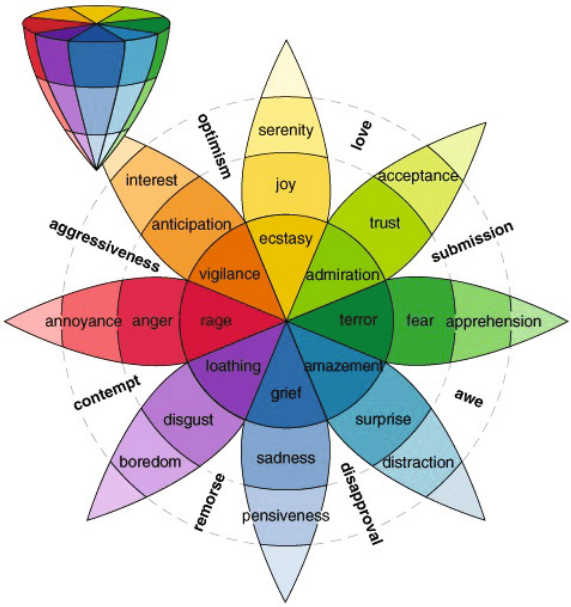

Fig. 3. Plutchik's primary and mixture emotions presented in a 2D wheel, and in a 3D cone [67].

emotional speech database (GES) [73], (2) Geneva vocal emotion expression stimulus set (GVEESS) [74] [1], and (3) Spanish emotional speech database (SES) [75]. [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 characteristics (i.e., pitch and energy [3]) have been calculated for all the samples of the databases [42] 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 4.1), features extraction (Section 4.2), and classification (Section 4.3), as indicated in Figure (4).

[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 [74].

[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.

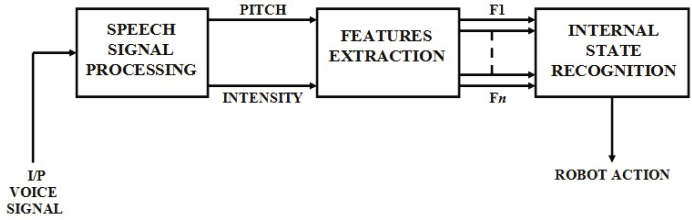

Fig. 4. Emotional state detection system

4.1 Speech Signal Processing

Talkin in [76] 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 processing. The signal conditioning process tries to clear away interfering signal components, such as any unnecessary noise by using low pass filtering, which removes any loss of periodicity in the voiced signal's spectrum at high frequencies, and by 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 [76] developed the traditional Normalized Cross Correlation (NCC) method [77,78] 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 normalized cross correlation (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 (5). On the other hand, voice signal's energy could be directly calculated from squaring the amplitude of signal's points.

4.2 Features Extraction

Rong et al., in [79] 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

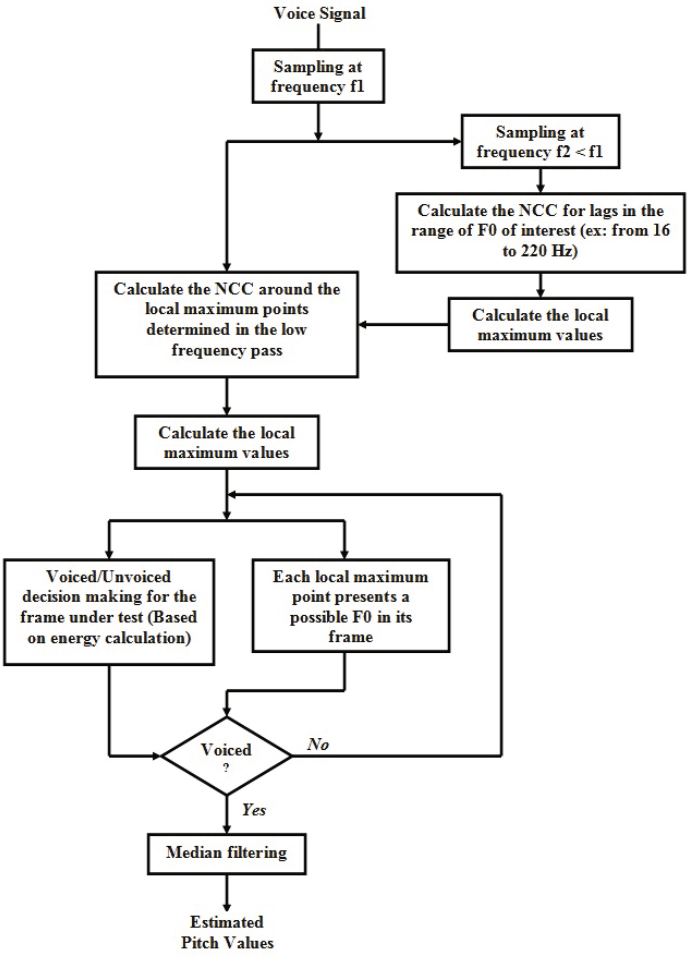

Fig. 5. Pitch tracking

recognition score. Relevant statistical measures of pitch and energy were calculated in order to create 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.

4.3 Classification

Voice samples were classified using the SVM algorithm with a quadratic kernel function [80,81], and the results were cross validated. Table (1) indicates the obtained recognition scores of 15 different emotions. The mean values of the

recognition scores indicated in Table (1), reflect the high precision of our classification system with respect to similar systems discussed in the literature. In [73], the mean value of the emotion recognition scores was 86.1%. Meanwhile, in [74], the mean value of the obtained scores was 60%. However, in [75], the mean value of the obtained scores was 85.9%.

Table 1. Recognition scores of different emotional states. Empty spaces are emotions not included in these databases.

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%

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 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 prescribed 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 lightly discussed in this work a future research orientation).

5 Subtractive Clustering

Subtractive Clustering [59] is a fast 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 (Equation 1):

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

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 reformulated as following (Equation 2, where r_b is a positive constant):

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

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) [59]. After calculating the reduced potential of all data points with respect to the last defined cluster center according to Equation (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 in [59] 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 3):

$$\frac{d_{min}}{r} + \frac{P_n^*}{P_1^*} \geq 1 \tag{3}$$

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 in [59], 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.

6 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 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 4) [82]:

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

Consequently, the total degree of membership of rule i with respect to the whole input vector could be defined as following (Equation 5):

$$\tau_i = \mu_{1i}(y_1) \times \mu_{2i}(y_2) \times \cdots \times \mu_{mi}(y_m) = \prod_{j=1}^{m} \mu_{ji}(y_j) \tag{5}$$

The previous model could be reformulated in terms of linguistic *if-then* fuzzy rule as following (Equation 6):

$$\begin{aligned} &\text{If } y_1 \text{ is } y_{1i}^* \text{ and } \cdots\cdots \text{ and } y_m \text{ is } y_{mi}^* \\ &\text{Then} \quad z_i^* = b_{0i} + b_{1i}y_1 + \cdots + b_{mi}y_m \end{aligned} \tag{6}$$

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 7):

$$z = \sum_{i=1}^{n} \frac{\tau_i z_i^*}{\sum_{l=1}^{n} \tau_l} = \sum_{i=1}^{n} \gamma_i z_i^* \tag{7}$$

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 fuzzy rules as following (Equation 8):

$$z = \chi\varphi \tag{8}$$

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 human emotion 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 9, where k is the number of data points within a cluster):

$$J = \sum_{d=1}^{k} (z_d - \chi\varphi_d)^2 \tag{9}$$

Equation (9) could be reformulated as following (Equation 10, where the matrices Z, η are functions in z_d and φ_d):

$$J = (Z - \chi\eta)^T (Z - \chi\eta) \tag{10}$$

The least square estimation of χ, could be finally defined as following (Equation 11):

$$\hat{\chi} = (\eta\eta^T)^{-1}\eta Z \tag{11}$$

A typical fuzzy modeling of a human's emotional state is illustrated in Figure (6), 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 (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 12, where α is the total number of the existing clusters):

$$Class = \arg\max_{p=1}^{\alpha}(z_p) \tag{12}$$

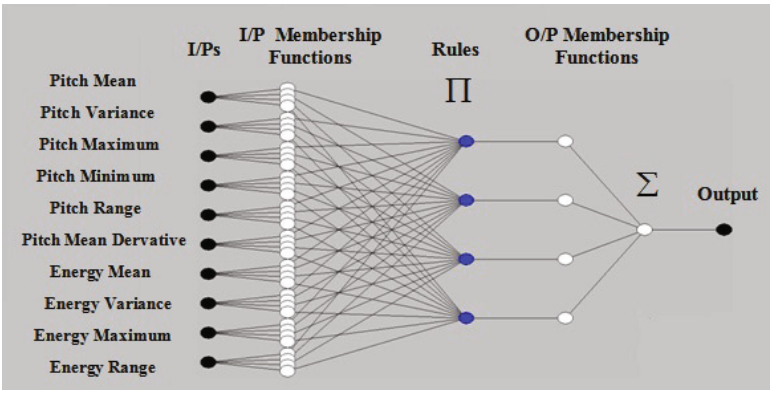

Fig. 6. TS fuzzy modeling of a human's emotion cluster

7 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 online incoming data [82] 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 7). When a new data element arrives, it gets attributed to one of the existing clusters according to Equation (12), which leads to one of the three scenarios below:

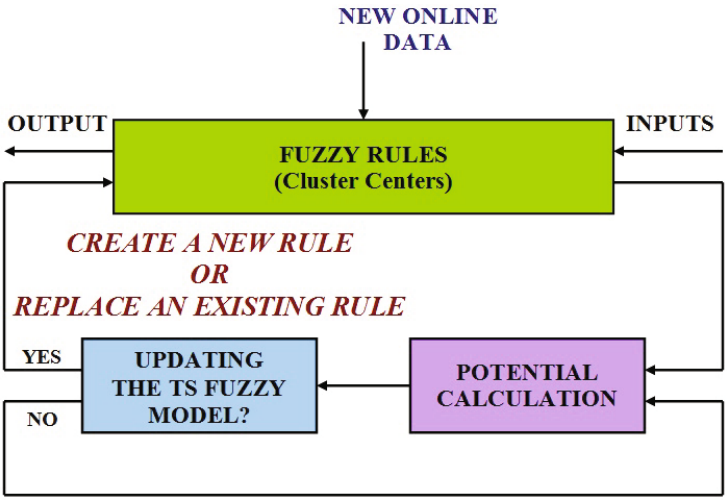

Fig. 7. TS fuzzy model updating whether by creating a new rule, or by replacing an existing rule.

7.1 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 (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 7). 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 (7 to 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). 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 [82].

7.2 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. Thereupon, 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). Consequently, a new cluster will be created (with an associated neutral action, until the robot gets able to synthesize 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.

7.3 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 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**

8 Results

The fuzzy classification system was trained on 7 emotions (i.e., anger, disgust, happiness, sadness, surprise, fear, neutral), and the results were cross validated (Table 2). The calculated scores are less than the previously obtained scores through the offline learning process using the SVM algorithm (Table 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.

Table 2. Recognition scores of the fuzzy system's training emotions

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%

The online test database included voice samples covering simple and complex emotions from the three databases referred to earlier (Section 4), 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 (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.

Table 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

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%

The part of the new data attributed to the existing clusters (Table 3), was assigned for the validation of Scenario 1 (Section 7). 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 3), was assigned for the validation of Scenario 2 (Section 7). 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, 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 (Figure 8) developed by Aldebaran Robotics [4] is available at: `http://www.ensta-paristech.fr/~tapus/HRIAA/media/videos/`. The video is composed of four scenes recognizing three emotions belonging to the existing clusters in the database (Figure 9), in addition to one new emotion not included 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).

Fig. 8. Test-bed: Nao robot

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 7) 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 algorithm would be able to detect the new emotion and to construct a new cluster.

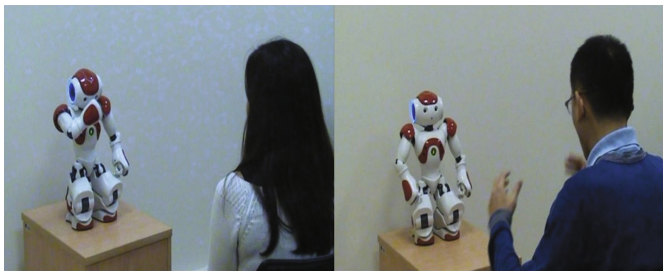

Fig. 9. Two participants are interacting with the robot. Each one expressed an emotion and the robot tried to recognize it. This recognition represents an action that the robot could generate corresponding to the expressed emotion.

9 Conclusion

This research illustrates the need for artificial cognitive functions that can allow the robot to understand and generate actions, in addition to understanding the emotional state of human so as to behave in a suitable way as human does. Mirror neurons and the Wernicke's area in human's brain play an important role in understanding the perceived multimodal actions in the surrounding environment. On the other hand, imitation and emulation are considered the most credible learning strategies in human's brain, because they provide the ability to generate an action according to the context. Moreover, the Broca's area in human's brain is believed to be responsible for speech production. The human cognitive model (Section 2), which interconnects between different brain areas and cognitive functions that organizes understanding and generating multimodal actions, represents a real challenge in front of the serious efforts towards creating a complete artificial cognitive model of similar functionality.

The major part of this research discusses an online learning approach for human's emotional states. 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.

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

[4] Nao is a 25 degrees of freedom robot equipped with two cameras, an inertial sensor, a sonar sensor, and many other sensors that allow it to perceive its surrounding with high precision and stability. `http://www.aldebaran-robotics.com/`

inconsistency in the context of interaction. Progressively, the robot's experience and awareness will increase, which helps it create autonomously a behavior from its own system by studying all the previous actions and interaction scenarios in order to propose autonomously new relevant actions. However, this last point is a future scope for this work.

Acknowledgments. This work is supported by the French National Research Agency (ANR) through Chaire d'Excellence program 2009 (Human-Robot Interaction for Assistive Applications). The project's website is accessible at: `http://www.ensta-paristech.fr/~tapus/HRIAA/`. This paper is an extension of our previous research [83], with more elaborated discussion and analysis.

References

1. Fogassi, L., Ferrari, P., Gesierich, B., Rozzi, S., Chersi, F., Rizzolatti, G.: Parietal lobe: From action organization to intention understanding. Science 308, 662–667 (2005)
2. Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G.: Action recognition in the premotor cortex. Brain 119, 593–609 (1996)
3. Schaffler, L., Luders, H., Dinner, D., Lesser, R., Chelune, G.: Comprehension deficits elicited by electrical stimulation of broca's area. Brain 116, 695–715 (1993)
4. Gazzola, V., Keysers, C.: 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 (2009)
5. Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., Rizzolatti, G.: Cortical mechanisms of human imitation. Science 286, 2526–2528 (1999)
6. Ramachandran, V.: Mirror neurons and imitation learning as the driving force behind "the great leap forward" in human evolution. Edge 69 (2000)
7. Rizzolatti, G., Fogassi, L., Gallese, V.: Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience 2, 661–670 (2001)
8. Rizzolatti, G., Arbib, M.: Language within our grasp. Trends in Neurosciences 21, 188–194 (1998)
9. Gallese, V., Goldman, A.: Mirror neurons and the simulation theory of mind reading. Trends in Cognitive Sciences 2, 493–500 (1998)
10. Ramachandran, V., Oberman, L.: Broken mirrors: A theory of autism. Scientific American 295, 62–69 (2006)
11. Ojemann, G., Ojemann, J., Lettich, E., Berger, M.: Cortical language localization in left, dominant hemisphere: An electrical stimulation mapping investigation in 117 patients. Neurosurgery 71, 316–326 (1989)
12. Whiten, A., Ham, R.: 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 (1992)
13. Whiten, A., Custance, D., Gomez, J., Teixidor, P., Bard, K.: Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). Comparative Psychology 110, 3–14 (1996)
14. Whiten, A.: Imitation of sequential and hierarchical structure in action: Experimental studies with children and chimpanzees. In: Cambridge, M.P. (ed.) Imitation in Animals and Artifacts, MA, USA, pp. 191–209 (2002)

15. Tomasello, M., Davis-Dasilva, M., Camak, L., Bard, K.: Observational learning of tool use by young chimpanzees and enculturated chimpanzees. Human Evolution 2, 175–183 (1987)
16. Tomasello, M.: Emulation learning and cultural learning. Behavior and Brain Science 21, 703–704 (1998)
17. Wood, D.: Social interaction as tutoring. In: Bornsten, M.H., Bruner, J. (eds.) Interaction in Human Development, Hillsdale, NJ, USA, pp. 59–80 (1989)
18. Whiten, A.: The scope of culture in chimpanzees, humans and ancestral apes. Philosophical Transactions of the Royal Society 366, 935–1187 (2011)
19. Galef, B.: The question of animal culture. Human Nature 3, 157–178 (1992)
20. Heyes, C.: Imitation, culture and cognition. Animal Behavior 46, 999–1010 (1993)
21. Tomasello, M., Savage-Rumbaugh, E., Kruger, A.: Imitative learning of actions on objects by children, chimpanzees and enculturated chimpanzees. Child Development 64, 1688–1705 (1993)
22. Buchsbaum, D., Griffiths, T., Gopnik, A., Baldwin, D.: 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 (2009)
23. Buchsbaum, D., Canini, K., Griffiths, T.: Segmenting and recognizing human action using low-level video. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society (2011)
24. Tani, J.: Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. Neural Networks 16, 11–23 (2003)
25. Tani, J., Ito, M., Sugita, Y.: Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using rnnpb. Neural Networks 17, 1273–1289 (2004)
26. Issar, S., Ward, W.: Cmu's robust spoken language understanding system. In: Proceedings of the 3rd European Conference on Speech Communication and Technology, EUROSPEECH (1993)
27. Bennacef, S., Bonnea-Maynard, H., Gauvain, J., Lamel, L., Minker, W.: A spoken language system for information retrieval. In: Proceedings of the 3rd International Conference on Spoken Language Processing, ICSLP (1994)
28. Miller, S., Bobrow, R., Schwartz, R., Ingria, R.: Statistical language processing using hidden understanding models. In: Proceedings of the Human Language Technology Workshop, NJ, USA (1994)
29. Levin, E., Pieraccini, R.: Concept-based spontaneous speech understanding system. In: Proceedings of the 4th European Conference on Speech Communication and Technology, EUROSPEECH (1995)
30. Goldberg, E., Driedger, N., Kittredge, R.: Using natural language processing to produce weather forecasts. IEEE Intelligent Systems and their Applications 9, 45–53 (1994)
31. Busemann, S.: Ten years after: An update on tg/2 (and friends). In: Proceedings of the European Natural Language Generation Workshop (2005)
32. Mcroy, S., Channarukul, S., Ali, S.: An augmented template-based approach to text realization. Natural Language Engineering 9, 381–420 (2003)
33. Bateman, A.: Enabling technology for multilingual natural language generation: The kmpl development. Natural Language Engineering 3, 15–55 (1997)
34. Lavoie, B., Rambow, O.: A fast and portable realizer for text generation. In: Proceedings of the 5th Conference on Applied Natural-Language Processing, ANLP (1997)

35. Gergely, G.: What should a robot learn from an infant? mechanisms of action interpretation and observational learning in infancy. Connection Science 15, 191–209 (2003)
36. Kozima, H., Nakagawa, C., Yano, H.: Emergence of imitation mediated by objects. In: Proceedings of the 2nd International Workshop on Epigenetic Robotics (2002)
37. Rudolph, M., Muhlig, M., Gienger, M., Bohme, H.: Learning the consequences of actions: Representing effects as feature changes. In: Proceedings of the International Symposium on Learning and Adaptive Behavior in Robotic System (2010)
38. Murray, I., Arnott, J.: Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. Journal of the Acoustical Society of America 93, 1097–1108 (1993)
39. Cahn, J.: Generating expression in synthesized speech. In Master's thesis, MIT Media Lab, USA (1990)
40. Roy, D., Pentland, A.: Automatic spoken affect analysis and classification. In: Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition, Vermont, USA (1996)
41. Slaney, M., McRoberts, G.: Baby ears: A recognition system for affective vocalizations. In: Proceedings of the 1998 International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, USA (1998)
42. Breazeal, C., Aryananda, L.: Recognition of affective communicative intent in robot-directed speech. Autonomous Robots Journal 12, 83–104 (2002)
43. Vogt, T., Andre, E.: Improving automatic emotion recognition from speech via gender differentiation. In: Proceedings of the Language Resources and Evaluation Conference, LREC 2006 (2006)
44. Voeffra, C.: Emotion-sensitive human-computer interaction (hci): State of the art. In: Seminar Emotion Recognition (2011), http://diuf.unifr.ch/main/diva/teaching/seminars/emotion-recognition
45. Pierre-Yves, O.: The production and recognition of emotions in speech: features and algorithms. Human-Computer Studies 59 (2003)
46. Jones, C., Deeming, A.: Affective human-robot interaction. In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction: From Theory to Applications, pp. 175–185 (2008)
47. Zadeh, L.: Fuzzy sets. Information and Control 8, 338–353 (1965)
48. Zadeh, L.: Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man, and Cybernetics 3, 28–44 (1973)
49. Mamdani, E., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies 7, 1–13 (1975)
50. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its application to modeling and control. IEEE Trans. on Systemsm Man, and Cybernetics 15, 116–132 (1985)
51. Sugeno, M.: Industrial applications of fuzzy control. Elsevier Science Pub. Co. (1985)
52. Bezdek, J.: Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York (1981)
53. Vapnik, V.: Statistical learning theory. In: Haykin, S. (ed.) Adaptive and Learning Systems. John Wiley and Sons (1998)
54. Dunn, J.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics 3, 32–57 (1973)
55. Gustafsson, D., Kessel, W.: Fuzzy clustering with a fuzzy covariance matrix. In: Proceedings of the IEEE CDC, San Diego, CA, USA, pp. 761–766 (1979)

56. Gath, I., Geva, A.: Unsupervised optimal fuzzy clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence 11, 773–781 (1989)
57. Yager, R., Filev, D.: Approximate clustering via the mountain method. In Technical Report MII 1305, Machine Intelligence Institute, Iona College, New Rochelle (1992)
58. Yager, R., Filev, D.: Learning of fuzzy rules by mountain clustering. In: Proceedings of SPIE Conference on Applications of Fuzzy Logic Technology, Boston, MA, pp. 246–254 (1993)
59. Chiu, S.: Fuzzy model identification based on cluster estimation. Journal of Intelligent and Fuzzy Systems 2, 267–278 (1994)
60. Searle, J.: Austin on locutionary and illocutionary acts. The Philosophical Review 77, 405–424 (1968)
61. Searle, J.: Speech acts: An essay in the philosophy of language. Cambridge University Press (1969)
62. Goldberg, L.: An alternative description of personality: The big-five factor structure. Personality and Social Psychology 59, 1216–1229 (1990)
63. Aly, A., 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 International Conference on Human-Robot Interaction, HRI (2013)
64. Summers-Stay, D., Teo, C., Yang, Y., Fermuller, C., Aloimonos, Y.: 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 (2012)
65. Pastra, K., Aloimonos, Y.: The minimalist grammar of action. Philosophical Transactions B 367, 103–117 (2012)
66. Izard, C.: Face of emotion. Appleton, New York (1971)
67. Plutchik, R.: The nature of emotions. University Press of America, Lanham (1991)
68. Ekman, P.: Emotion in the human face: Guidelines for research and an integration of findings. Pergamon Press, New York (1972)
69. Ekman, P., Friesen, W., Ellsworth, P.: What emotion categories or dimensions can observers judge from facial behavior? In: Ekman, P. (ed.) Emotion in the Human Face. Cambridge University Press, New York (1982)
70. Izard, C.: Human emotions. Plenum Press, New York (1977)
71. Tomkins, S.: Affect theory. In: Scherer, K., Ekman, P. (eds.) Approaches to Emotion, pp. 163–195. Erlbaum, Hillsdale (1984)
72. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20, 273–297 (1995)
73. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Proc. of Interspeech, Germany (2005), `http://database.syntheticspeech.de`
74. Banse, R., Scherer, K.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70, 614–636 (1996)
75. Montero, J., Gutierrez-Arriola, J., Palazuelos, S., Enriquez, E., Aguilera, S., Pardo, J.: Emotional speech synthesis: from speech database to tts. In: Proceedings of the International Conference on Spoken Language Processing 1998, pp. 923–925 (1998)
76. Talkin, D.: A robust algorithm for pitch tracking. In: Kleijn, W.B., Paliwal, K. (eds.) Speech Coding and Synthesis, pp. 497–518. Elsevier (1995)
77. Sondhi, M.: New methods of pitch extraction. IEEE Trans. Audio and Electroacoustics 16, 262–266 (1968)

78. Rabiner, L., Atal, B., Sambur, M.: Lpc prediction error: Analysis of its variation with the position of the analysis frame. IEEE Trans. on Systems Man, and Cybernetics 25, 434–442 (1977)
79. Rong, J., Li, G., Chen, Y.P.: Acoustic feature selection for automatic emotion recognition from speech. Information Processing and Management 45, 315–328 (2008)
80. Cristianini, N., Shawe-Taylor, J.: Introduction to support vector machines. Cambridge University Press (2000)
81. Platt, J.: Fast training of support vector machines using sequential mininal optimization. In Microsoft Research Technical Report MSR-TR-98-14 (1998)
82. Angelov, P.: Evolving rule-based models: A tool for design of flexible adaptive systems. STUDFUZZ, vol. 92. Springer, Heidelberg (2002)
83. Aly, A., Tapus, A.: Towards an online real time fuzzy modeling for human internal states detection. In: Proceedings of the 12th IEEE International Conference on Control, Automation, Robotics and Vision, ICARCV (2012)