A Socially-Intelligent Multi-Robot Service Team for In-Home Monitoring

Konstantinos Drossos*, Andreas Floros*, Stelios Potirakis†, Nikolas-Alexander Tatlas† and Gurkan Tuna‡
*Audiovisual Signal Processing Lab., Deptartment of Audiovisual Arts, Ionian University, Corfu, Greece
Email: {kdrosos,floros}@ionio.gr
†Department of Electronics, Technological Educational Institute of Piraeus, Egaleo, Greece
Email: {spoti,ntatlas}@teipir.gr
‡Department of Computer Programming, Trakya University, Edirne, Turkey
Email: gurkantuna@trakya.edu.tr

***Abstract*—The objective of this study is to develop a socially-intelligent service team comprised of multiple robots with sophisticated sonic interaction capabilities that aims to transparently collaborate towards efficient and robust monitoring by close interaction.In the distributed scenario proposed in this study, the robots share any acoustic data extracted from the environment and act in- sync with the events occurring in their living environment in order to provide potential means for efficient monitoring and decision-making within a typical home environment. Though, each robot acts as an individual recognizer using a novel emotionally-enriched word recognition system, the final decision is social in nature and is followed by all. Moreover, the social decision stage triggers actions that are algorithmically distributed among the robots'A population and enhances the overall approach with the potential advantages of the team work within specific communities through collaboration.**

I. Introduction

Growing numbers of elderly people have led the research community to find ways to aid the elderly in doing several everyday tasks, which are usually hard to accomplish by them. Considering the high cost raised for employing caretakers, robots may be preferred for this kind of tasks. Since the robots behavior is algorithmic in nature, they do not get bored and they do not get tired. Hence, they can be considered as attractive alternatives for caretaking services. However, one can note an increasing research interest in the use of robots for human caretaking environments [1]. The implementation of autonomous caretaking services is severely restricted by several challenges, which significantly limit the interaction of the robots and the human. For example, a highly complicated and time-critical interaction service is needed to accomplish efficient communication in real-time between the machine and the human [2]. This communication can be visual (i.e. through gestures) or aural (performed through common speech). Therefore, a special attention should be given to improve the communication skills of the robots involved.

Focusing on verbal communication only, apart from the content of the speech itself, an efficient communication framework should also include affective conditioning. It is well-known that speech conveys emotions [3]. Human listeners are able to perceive them and derive a combinative, emotionally-enriched information outcome [4]. This is a significant aspect that may enhance the human-machine interaction framework, provide additional means for securing the correctness of the decisions that a machine should take and allow for a human-robot interaction that tends to be a social relation [5]. Apart from the above communication capabilities enhancements, autonomous multi-robot service teams are still an unexplored area. Although multi-robot security teams have been proposed, e.g., [6], to the best of our knowledge, there exists no work on multi-robot service teams designed for caretaking services. To address this need, in this paper, we propose a multi-robot service team with sophisticated interaction capabilities. More specifically, in the proposed scenario, the robots share any acoustic data extracted from their environment to collaborate towards efficient and robust servicing by close interaction. In a distributed fashion, the robots act in-sync with the acoustic events occurring in the environment. Each one of them performs an assessment of the meaning and the emotional content of the speech, as well as of the sound characteristics. The final decision, on which is the appropriate action and which robot should perform it, is taken centrally by the higher ranked robot which has also available other information like the locations of both robots and human/s, special abilities of each robot (if any), etc. This achieved by three sequentially employed linguistic (Mamdani) fuzzy inference systems [7] and references therein. For the purposes of this work, we have considered equivalent multi-functional robots, instead of using several robots with different capabilities, while the acoustic events considered are limited to emotionally-enriched discrete speech events, i.e. words or verbal commands. Additional sound events can be considered (for example glass-breaking or human falling sounds) and potentially combined with the outcome of the affective sound interaction path. However, this investigation is considered to be out of the scope of this work, which aims to primarily investigate the potential impact and efficiency of advanced socially distributed intelligence provided though modern means of emotionally-enhanced acoustic interaction. The remainder of this paper is organized as follows. I Section II, related work on socially-intelligent service robots is summarized. Section III presents the overall system architecture. Overview of simulation results are presented in Section IV. Finally, the paper is concluded in Section V.

II. Related Work

The increasing interest on service robots has led the academia to study on designing robots having communication

skills. For the service robots to be used in home environments to care for the elderly, specific communication requirements should be met to enable the robots be socially-intelligent, one of the most important requirements of service robots. This section discusses both the interaction systems proposed for single-robot and multi-robot teams and the distributed control systems in the literature. Situational awareness is one of the key requirements of interactive service robots. Luo and Chang in [8] explain the details of two different robots called Chung-Cheng I and Security Warrior. Both robots utilize multi-sensor fusion processes for the detection and recognition of people. Basically, multi-sensor fusion and integration is the synergistic combination of data from multiple sensors to achieve inferences not feasible from each individual sensor operating separately [8]. Multi-sensor fusion brings several advantages and is used in many military and non-military applications. One of the best applications of multi-sensor fusion is simultaneous localization and mapping (SLAM) processes. SLAM is the key of autonomous robot systems and enables the robots to localize themselves and at the same time map the environment. In [9], the authors propose an intelligent service robot which creates an information-enriched map constructed by the environment geometry from a laser range finder and a camera. A similar work which explains the details of a navigation framework for multiple autonomous robots is explained in [10]. The proposed navigation framework was implemented in [10] on real robots called "Robox" and the robots operated during the Expo.02 exhibition. Different from these studies, human-robot interaction (HRI) skills are needed for most service robots. The work in [11] proposes a catering service robot to be used in restaurants. The robot is integrated with multimodal human computer interaction techniques and is supposed to take the place of restaurant staffs.

HRI is very important for service robots and enables them to understand the requirements of their users and identify where the users are. For the humans, voice is the most straightforward way to communicate. Therefore, automatic speech recognition systems should be integrated into service robots designed for home uses. A combined sound source localization and stereo vision system for service robots is proposed in [12]. A further step towards a sophisticated socially-intelligent robot is the ability to recognize emotional states of the human. The work proposed in [13] is a real-time emotion recognition system combined with a complicated recognition engine which recognizes facial expressions and categorizes them into one of seven different emotional states: happiness, sadness, fear,disgust, anger, surprise, and neutrality. In some scenarios, such as the one given in [14], other assistive technologies beyond recognition systems may be required. The work in [14] proposes a robotic agent that understands users' wishes and gives their possible answers on a social network platform by utilizing natural language processing (NLP) and metadata analysis. KSERA, a system consisting of an intelligent home environment, which incorporates smart home functionalities, a sensor-data-based inference module capable of detecting critical conditions and alarm raising functionality, is proposed in [15]. KSERA is complemented by a socially assistive robot. The robot is used as primary user interface for interaction with the users. The usability of the system was verified by the trial participants in Austria and Israel. Brian 2.0, a socially assistive robot to be used as a therapeutic aid designed to maintain, and improve the residual social and cognitive functioning in people with dementia, is proposed in [16]. Brian 2.0 is able to engage the people with dementia in the activity by means of task assistance,encouragement, reinforcement, and celebration. The work in [17] addresses several aspects of a robot called "CompanionAble" which was developed as part of the European FP7 project. CompanionAble is a socially assistive robot for elderly people with mild cognitive impairment (MCI) living alone at home.A similar system is proposed in [18]. This work aims to develop a socially intelligent robot, which may support diabetic patients to cope with their illness better by providing them guidelines. Different from the abovementioned socially assistive robots, a good discussion of service robots with manipulation skills to help the disabled/limited people is given in [19].

In this work a hierarchically organized team of service robots is proposed. Although all robots have equivalent capabilities, there is a hierarchical organization. The higher ranked robot takes the decisions. In case of failure of the higher ranked robot, the second in hierarchy takes over and so on. The action that should be taken in each situation by the team of robots is primarily determined by the answer to two crucial questions: (i) "what should be done?", and(ii) "who should do it?". The answer to both questions is given by a two-level distributed decision making system primarily based on sound characteristics, speech meaning and emotion as well as localization data.

III. System Architecture

The proposed system consists of a team of multifunctional robots of the same capabilities, and acting co-operatively, and two modules responsible for process the information presented to the robots. Each one of the robots is equipped with two acoustic sensors, a local signal processing unit, a localization system, a wireless transceiver for the communication among them, and a GSM modem for SMS messages. During the set-up of the robots' team a hierarchy among the team members is randomly set, since all robots are in principle equivalent. The coordination of their team is a task of the higher ranked available robot which emits periodically a beacon signal to denote its presence. If the latter signal is not emitted as expected for two periods, then the second in the hierarchy robot takes over the coordination tasks, and so on. The identification of the positions of both human/s and robots is performed by a trilateration-based localization scheme with several smart wireless sensor nodes located at predetermined fixed positions and a group of mobile robots. In following subsection are presented in detail the above mentioned modules, i.e. Emotionally-Enriched Word Recognition (EEWR) System and the Socially-Enriched Decision-making (SEDM) System.

A. The Emotionally-Enriched Word Recognition System

The EEWR system is able to recognize a finite set of speaker independent spoken words, enhanced with the ability to categorize them based on the human speaker stress level. The latter represents a significant affective component for applications targeted to automated home monitoring, since it may provide efficient means for prioritization of the recognized verbal objects. In general, affective speech recognition is based on the extraction of voice-signal technical features, their direct

Table I. EMOTION PROFILES FOR ANGER, FEAR, SADNESS AND HAPPINESS COMPARED TO NORMAL SPEECH

Emotions	Acoustic Cue		
	F_0	S_R	E
Anger	Increase	Increase	Increase
Happiness	Increase	Decrease	Increase
Sadness	Decrease	Decrease	Decrease
Fear	Increase	Increase	Increase

Table II. THE SET OF WORDS EMPLOYED FOR TRAINING

Word	Pronunciation	Translation	Word Index
Φωτιά	fɔ.ˊtça	Fire	1
Φώτα	ˊfɔ.ta	Lights	2
Σεισμός	si.ˊzmɔs	Earthquake	3
Καλά	ka.ˊla	Well/OK	4
Αλλά	a.ˊla	But	5
Κλέφτης	ˊkleftis	Thief	6
Καλημέρα	ka.li.ˊmɛ.ɾa	Goodmorning	7

comparison to appropriately defined thresholds, derived by categorization algorithms applied on a ground-truth data set, and the mapping of results to emotions employing an affective model. Many of such models already exist in the literature, ranging from discrete up to dimensional ones [20]. Typical emotion-recognition algorithms usually employ the voice fundamental frequency F_0 (i.e. the pitch) as the fundamental acoustic cue [21], while additional cues may be considered in parallel, such as the speech rate S_R (or the speech tempo) [22] and the instantaneous voice energy E [23]. Different ranges of these cues values are related to different emotions, forming the so-called acoustic/emotions' profiles. Table I illustrates some indicative profiles derived from the literature for typical emotions, such as anger, fear, sadness and happiness, compared to emotionally neutral (or normal) speech.

Many existing research works employ the Arousal-Valence (AV) affective model. Since stress is not directly included in the discrete emotion set, one can consider the same acoustic profile for it, provided that stress is a common component of both aforementioned emotions, especially under emergency situations particularly considered in this work. The stress recognition process employed in this work was initially introduced by the work carried out in [24]. We hereby provide a brief overview of its functional characteristics and architecture. For the interested reader, a detailed analysis is included in [24]. It particularly consists of three modules: a) the Voice Activity Detector (VAD), b) the Voice Keyword Recognizer (VKR) and c) the Voice Stress Classifier (VSC). The VAD module provides voice-active or voice-inactive estimates during the real-time capturing of the environmental sound. Voice-active recording periods are fed to the VKR subsystem which indicates the recognised word. During the VKR training session, the results provided were stored in the system internal storage (flash memory). The small size of the available memory in the embedded, DSP-free platform employed represents a major limitation for the selection of the training data set volume. Thus, in order to allow the increment of the recognized words number, only normal (unstressed) words were used as training data. The VSC module is responsible for providing stress or no-stress estimates for the words recognized by the VKR. For this reason, it calculates the acoustic cues presented in Table I using the waveforms of a recognized word. If all these values exceed the corresponding thresholds defined during the VSC training period, then there is an indication that the speaker experiences stress. The emotionally-enriched word recognition system was trained and tested for a set of seven Greek words, four of which have similar pronunciation in pairs. The specific words along with their pronunciation according to the International Phonetic Alphabet (IPA) is provided in Table II. Note that, the pronunciation of "Φωτιά" is similar to "Φώτα", and "Καλά" is similar to "Αλλά". The recognized word is coded as an integer index corresponding to the appropriate word from the data corpus as illustrated in Table II. Accordingly, the indication of stress is coded as a simple Boolean variable, i.e. true for stress indication and false otherwise. The distance of the robot and the percentage of the recognized word are both floating-point numbers. The former is the distance in meters, whereas the latter is calculated as the percentage of the z neighbors agreed on the recognized word out of the total k ones, with $z \leq k$. All the above information is finally transmitted to the SEDM system, executed on the higher ranked robot.

B. The Socially-Enriched Decision-making System

The SEDM system proposed in this work is a two-level system. The first level runs on each one of the robot team members, providing an assessment of the speech meaning and emotional content, as well as a rough estimation of the reliability of the locally reached decision on the specific speech-related information. In parallel, each one of the robots measures/assesses other parameters like the SPL at its location and its position in the house. The second level runs centrally on the higher ranked robot, based on the aforementioned information transmitted by the individual robots of the team. The final decision process takes into account the locally reached decisions and measurements of all team members, tries to reach a decision that will be followed by all team members and is implemented in three steps.

The first step is implemented as a six input, one output, Mamdani type FIS, using thirteen rules, and is independently running for each one of the robot team members. The employed inputs, all coming from the specific robot team member under evaluation, are: (i) the locally reached rough estimation of the reliability of the decision on the meaning and emotion for the left "ear", (ii) the corresponding reliability for the right "ear", (iii) the coincidence, or not, of the locally decided meanings for the left and the right ear, (iv) the coincidence, or not, of the locally decided emotions for the left and the right ear, (v) the robot distance from the sound source, and (vi) the SPL measured at the position of the specific robot. Once the credibility of all robots has been evaluated, the credibility scores are passed to the second step along with the meaning and the emotion inferred locally for each one of the robots. Therefore, the second step is implemented as a Mamdani type FIS, having three inputs for each robot, i.e., the number of inputs equals the product of the number of robots by three, and produces two outputs, the meaning and the emotion; unfortunately, the number of rules it uses is generally large, as the lower limit of the possible number of rules would be equal to the product of the number of robots by the number of words to be recognized plus one (the "unrecognized"). In this work, a set of 62 rules was used. Figure 1 depicts the membership functions corresponding to the considered cases for the output variable "meaning" of the FIS in the second step. Finally, the final step takes into account the finally inferred

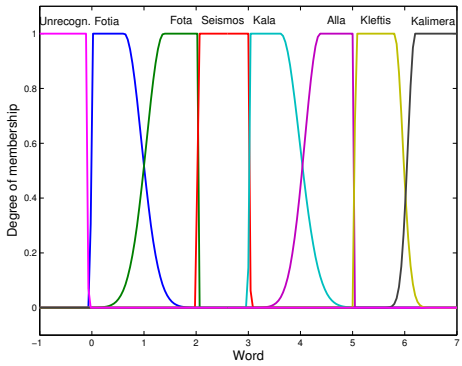

Figure 1. The membership functions corresponding to the recognized words

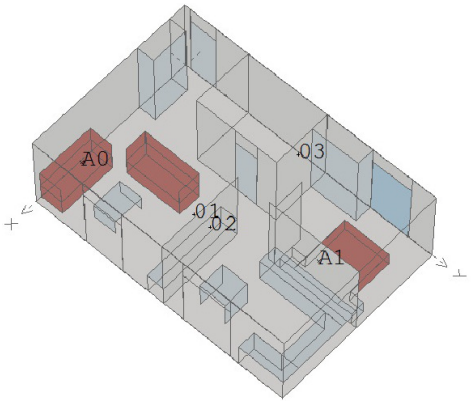

Figure 2. 3D Geometrical Model of the employed room

Table III. ACCURACY SCORES FOR THE VKR AND VSC SUBSYSTEMS FOR THE LEFT AND RIGHT BINAURAL CHANNEL

Stressed/Unstressed Words	VKR	VSC
Left binaural channel		
Unstressed (mean)	80.95%	100.00%
Stressed (mean)	69.04%	38.88%
Total mean	74.99%	69.44%
Left binaural channel		
Unstressed (mean)	95.04%	92.06%
Stressed (mean)	69.04%	54.16%
Total mean	82.04%	73.11%

meaning and emotion from the previous step along with the distance between each robot and the sound source, and the SPL measured at the position of each robot. This step reaches two decisions: (a) what is the action to be done and (b) which robot should do it. Therefore, the final step is implemented as a Mamdani type FIS, having as inputs the two outputs of the previous step plus two more inputs for each robot, and produces two outputs; the number of rules depends on the different actions that the robot will be programmed to perform. In our investigation, we considered that the number of the different possible actions equals the product of the different meanings to be recognized by the number of the different possible emotions plus one, the last action being a request to the human for repetition of the spoken word due to unresolved meaning/emotion.

IV. PERFORMANCE EVALUATION

Two test scenarios, corresponding to different relative positions of one human and tree equivalent multifunctional robots were employed for the evaluation of the EEWR and SEDM systems.. Specifically, two different sound source (human position) cases were considered for the same receiver (robot) positions inside a typical home environment. As an acoustical setup a typical three room (living room, kitchen and bedroom) apartment was considered, with an approximate total surface of $54m^2$ and surfaces with generally reflective materials: for example tiled floor, painted concrete walls and windows without curtains are used throughout the apartment, leading to an acoustically challenging environment. Two speaker positions were defined, shown as "$A0$" (living room) and "$A1$" (bedroom), while three robots are positioned shown as "01", "02", "03" in Figure 2. The voice signal emitted by the human is estimated at the positions of the robots for each one of their "ears" for the above two scenarios. Note that, robot positions "01" and "02" were deliberately positioned in almost identical distance from the human in both test scenarios (human positions "A0", and "A1"), and all three robot positions were put in very similar distances from "A1", to test whether the decision making system can efficiently exploit the acoustics-related information that can be acquired by the robots in order to decide which robot should act in each test case. A detailed geometrical/surface model for the apartment was created and simulated using a well-established software platform [25]. Figure 2 depicts the three-dimensional geometrical model, obtained from the simulation tool. Following the acoustic simulation for each speaker, two audio streams (left and right ear) per robot position were created through the auralization module, representing the actual voice signal captured by each robot. In the following subsections the functionality tests are presented along with the obtained results giving an indication of the efficacy of the proposed approach under realistic conditions.

A. Voice keyword and stress recognition

The performance of the VKR and VSC subsystems was evaluated utilizing a test data set recorded by two male speakers. All seven words, as presented in Table II, were recorded for both stressed and un-stressed cases. For this reason, both speaker were trained actors that can verbally express emotions in a fully controllable fashion. The technical parameters of the voice recording process were: 16 kHZ sampling frequency, 1 channel and 16 bits sample length. In addition, the recorded material served as input to the auralization module of the CATT-Acoustic software used for the acoustic simulations [25]. The output was a binaural version of each input, which was further used as test data for the VKR and VSC subsystems, corresponding to the input acoustic signal for each of the robots' ears within the home environment considered in this work. The results of the recognition process implemented by the aforementioned subsystems are summarized in Table III for the left and right binaural channels.

As it can be observed from the above summary of results, the mean word recognition accuracy for both stressed and unstressed words is 82.04% and 74.99%, for the right and left channel respectively. Also, the obtained stress identification results derived from the VSC subsystem are 73.11% and 69.44% (again for the right and left binaural channels). These accuracy

Table IV. DECISIONS REACHED BY THE SOCIALLY-ENRICHED DECISION MAKING SYSTEM. COLOR SHADINGS HAVE THE FOLLOWING MEANING: YELLOW INDICATES "UNRECOGNIZED", GREEN INDICATES INCORRECT EMOTION BUT CORRECT WORD AND BLUE INDICATES THAT BOTH WORD AND EMOTION INCORRECTLY INFERRED.

Word Index	Emotion	Req. Action	Decision			
			$A0$ Position		$A1$ Position	
			Action	Robot	Action	Robot
1	no-stress	1	1	01	1	03
	stress	2	2	01	2	03
2	no-stress	3	3	01	3	03
	stress	4	4	01	4	03
3	no-stress	5	5	01	5	03
	stress	6	6	01	6	03
4	no-stress	7	7	01	-1	03
	stress	8	8	01	1	03
5	no-stress	9	2	01	2	03
	stress	10	-1	02	10	03
6	no-stress	11	4	01	4	03
	stress	12	12	01	12	03
7	no-stress	13	6	01	13	03
	stress	14	12	01	7	03

values are clearly degraded compared to the VKR and VSC identification efficiency reported in [24], where all accuracy measurements where derived for anechoic sound recordings. Clearly, the close room acoustic properties represent a significant parameter for the correct word recognition, while it is found not to significantly affect the stress identification accuracy.

B. Decision making system

The decisions obtained by the SEDM system for the 28 different test cases (7 words by 2 possible emotions by 2 considered source-receivers scenarios) are presented in Table IV. For each test scenario (human position) 15 different possible "actions" were considered, corresponding to the 14 possible word-emotion combinations. From these results one can see that the decision making system reaches the correct decision for all, but one, cases concerning the robot that should perform the action, since it selects the robot that is both in near distance and in the same room, leading to an overall score of 96.43% correct decision. Note that, as already pointed out, robot positions "01" and "02" are in almost identical distance from the human in both test scenarios (human positions "A0", and "A1"). From the results presented in Table IV we can see that the decision making system reaches the correct decision in the vast majority of the non-stress cases. Specifically, all non-stress actions (100%) were recognized for the first test scenario (human placed at position "A0"), while only one of the non-stress actions was unrecognized for the second test scenario (human placed at position "A1"), leading to a success rate of 85.71% for this scenario. Importantly, this means that both the word and the emotion were correctly inferred, even though, as already mentioned, the pronunciation of "Φωτιά" is similar to "Φώτα", and "Καλά" is similar to "Αλλά". If one would present single-figure scores for the non-stressed actions, then 92.86% of the actions, the words and the emotions were correctly inferred.

On the other hand, concerning the stressed words, although the word part related to each action was correctly assessed in most cases, yielding a success rate of 71.43% when human is placed at position "A0" and 100% for the "A1" position, the emotion was not successfully assessed. Specifically, only

Table V. WORD/EMOTION RELATED INPUTS AND OUTPUTS OF THE SECOND STEP OF THE SEDM SYSTEM FOR THE FIRST AND SECOND TEST SCENARIOS (HUMAN AT POSITION "A0" AND "A1" RESPECTIVELY). "R" STANDS FOR "ROBOT", "N" AND "S" ARE FOR NO-STRESS AND STRESS REPSECTIVELY. "W_i" IS THE WORD INDEX AND "E" IS FOR "EMOTION".

Emitted Information		Locally Recognized Information						Final Decision	
		E			W_i			Inferred	
W_i	E	R.1	R.2	R.3	R.1	R.2	R.3	W_i	E
Human at "A0"									
1	N	0	0	0	1	1	1	1	N
2	N	0	0	0	2	2	2	2	N
3	N	0	0	0	3	3	3	3	N
4	N	0	0	0	4	4	4	4	N
5	N	0	0	0	5	5	5	5	N
6	N	0	0	0	6	6	6	6	N
7	N	0	0	0	7	7	7	7	N
1	S	0	1	1	1	1	1	1	S
2	S	1	0	0	2	2	2	2	N
3	S	0	-1	-1	3	3	-1	3	-1
4	S	1	0	0	4	1	4	4	N
5	S	1	1	1	5	5	2	5	S
6	S	0	0	0	6	6	6	6	N
7	S	1	1	1	1	5	5	1	S
Human at "A1"									
1	N	0	0	0	1	1	1	1	N
2	N	1	0	0	2	2	2	2	N
3	N	0	0	0	3	3	3	3	N
4	N	0	0	0	4	4	4	4	N
5	N	0	0	0	5	5	5	5	N
6	N	0	0	0	6	6	6	6	N
7	N	-1	0	0	-1	4	7	5	N
1	S	1	0	0	1	1	1	1	N
2	S	1	0	0	2	2	2	2	N
3	S	1	1	1	3	-1	3	3	S
4	S	0	0	0	4	4	4	4	N
5	S	1	1	1	5	5	5	5	S
6	S	0	1	1	6	6	6	6	S
7	S	1	0	0	7	7	7	5	N

for the 42.86% of the cases the emotion was successfully inferred for either of the two considered human positions. Unfortunately, the incorrect assessment of the emotion lead to a quite limited mean score (35.71%) for the correct decision on the required action, although a quite high mean score (85.71%) was achieved on the word inference. This means that in the cases for which the required action was not correctly decided this was due to incorrect emotion inference. Table V refers to the inputs and outputs of the second ("Meaning/Emotion") step of the decision making system.

V. CONCLUSION

In this work we propose an integrated approach of a socially-intelligent multi-robot service for human caretaking in home environments based on emotionally-enriched, speech-based interaction. The overall approach incorporates a combination of algorithms for speaker-independent keyword recognition and combined stress (or no-stress) identification. Keyword identification is performed over a limited set of words that are used in everyday speech but are also associated to emergency situations (such as fire, etc). Within this framework, stress identification may significantly affect the final decision that should be taken in the presence of the speaker stress conditions, primarily in terms of prioritization of the triggered actions that should be taken by the home monitoring system. Although the emotionally-enriched voice keyword recognition strategy achieves high accuracy efficiency in anechoic conditions, the final results obtained in typical home environments are found to be significantly degraded, especially for the word recognition task. In order to improve this performance, in this paper we take advantage of the randomly placed robots that act as

binaural, emotionally enriched word recognizers, aiming to provide the data derived by them to a higher level word/stress recognition system that is responsible for producing the final decision, including the selection of the robot that should act as a response to the emergency condition.

This decision making system proposed hereby is executed on the robot marked with the higher rank among the team members. All the necessary information exchange between the robots is performed over a typical wi-fi data network. In all test/simulation cases considered in this work, the proposed decision-making system was proved quite successful in selecting which robot should perform the action, while the word part of the selected action was also quite successfully inferred, increasing the mean success rate of the word recognition achieved by the individual robots. The same holds for the inference of the emotion in the case of "no-stress". Therefore the selected action for the test cases with no-stress emotion was also quite successfully decided. On the other hand, the "stress" emotion was not successfully enough inferred, leading to a consequent low score regarding the correct action decision for the corresponding test cases. Future extensions of this work may include the continuous robot-based evaluation of the acoustic properties of their surrounding environment that may be associated with additional weighting functions employed by the centralized decision-making system, towards fully-efficient and safe home robot caretaking environments.

REFERENCES

[1] A. Wykowska and A. Schubŏ, "Perception and action as two sides of the same coin. a review of the importance of action-perception links in humans for social robot design and research," *International Journal of Social Robotics*, vol. 4, no. 1, pp. 5–14, 2012. [Online]. Available: http://dx.doi.org/10.1007/s12369-011-0127-6

[2] M. B. Ammar, M. Neji, A. M. Alimi, and G. Gouardéres, "The affective tutoring system," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3013–3023, 2010.

[3] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," in *The new handbook of Methods in Nonverbal Behavior Research*, J. A. Harrigan, R. Rosenthal, and K. R. Scherer, Eds. New York, U.S.A.: Oxford University Press, 2008, ch. 3, pp. 65–136.

[4] T. Johnstone, C. M. van Reekum, T. R. Oakes, and R. J. Davidson, "The voice of emotion: an fmri study of neural responses to angry and happy vocal expressions," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 3, pp. 242–249, 2006.

[5] C. L. Lisetti, S. M. Brown, K. Alvarez, and A. H. Marpaung, "A social informatics approach to human-robot interaction with a service social robot," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 2, pp. 195–209, 2004.

[6] M. Lee, M. Tarokh, and M. Cross, "Fuzzy logic decision making for multi-robot security systems," *Artificial Intelligence Review*, vol. 34, no. 2, pp. 177–194, 2010. [Online]. Available: http://dx.doi.org/10.1007/s10462-010-9168-8

[7] S. M. Potirakis, G. E. Alexakis, M. C. Tsilis, and P. J. Xenitidis, "Time-domain nonlinear modeling of practical electroacoustic transducers," *Journal of Audio Engeering Society*, vol. 47, no. 6, pp. 447–468, 1999.

[8] R. C. Luo and C. C. Chang, "Multisensor fusion and integration aspects of mechatronics," *IEEE Industrial Electronics Magazine*, vol. 4, no. 2, pp. 20–27, 2010.

[9] R. C. Luo and C. C. Lai, "Enriched indoor map construction based on multi-sensor fusion approach for intelligent service robot," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 4, pp. 3135–3145, 2012.

[10] K. O. Arras, R. Philippsen, N. Tomatis, M. Battista, M. Schilt, and R. Siegwart, "A navigation framework for multiple mobile robots and its application at the expo.02 exhibition," in *Proceedings of the IEEE International Conference on Robotics and Automation, 2003, ICRA '03*, 2003, pp. 1992–1999.

[11] C. Chen, Q. Gao, Z. Song, O. Liping, and X. Wu, "Catering service robot," in *Proceedings of the 2010 8th World Congress on Intelligent Control and Automation (WCICA)*, 2010, pp. 599–604.

[12] R. C. Luo, W. H. Cheng, and C. H. Huang, "Combined 2-d sound source localization with stereo vision for intelligent human-robot interaction of service robot," in *Proceedings of the 2009 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2009, pp. 24–29.

[13] R. Alazrai and C. S. G. Lee, "Real-time emotion identification for socially intelligent robots," in *Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA)*, 2012, pp. 4106–4111.

[14] H. J. Min and J. C. Park, "Make your wishes to 'genie in the lamp': Physical push with a socially intelligent robot," in *Proceedings of the 6th international conference on Human-robot interaction, HRI '11*, 2011, pp. 203–204.

[15] K. Werner, J. Oberzaucher, and F. Werner, "Evaluation of human robot interaction factors of a socially assistive robot together with older people," in *Proceedings of the 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, 2012, pp. 455–460.

[16] J. Chan and G. Nejat, "Promoting engagement in cognitively stimulating activities using an intelligent socially assistive robot," in *Proceedings of the 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, 2010, pp. 533–538.

[17] H. M. Gross, H. M. Gross, C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley, C. Martin, T. Langner, and M. Merten, "Progress in developing a socially assistive mobile home robot companion for the elderly with mild cognitive impairment," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011, pp. 2430–2437.

[18] R. Looije, F. Cnossen, and M. A. Neerincx, "Incorporating guidelines for health assistance into a socially intelligent robot," in *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006*, 2006, pp. 515–520.

[19] B. Graf and H. Staab, "Service robots and automation for the disabled/limited," in *Springer Handbook of Automation*, S. Y. Nof, Ed. New York, NY: Springer-Verlag Berlin Heidelberg, 1988, ch. 84, pp. 1485–1502.

[20] K. Sun, J. Yu, Y. Huang, and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo, 2009. ICME 2009*, 2009, pp. 566–569.

[21] S. Giripunje and N. Bawane, "Anfis based emotions recognition in speech," in *Knowledge-Based Intelligent Information and Engineering Systems*, ser. Lecture Notes in Computer Science, B. Apolloni, R. J. Howlett, and L. Jain, Eds. New York, NY: Springer-Verlag Berlin Heidelberg, 2007, vol. 4692, ch. Part I, pp. 77–84.

[22] K. R. Scherer, "Vocal affect signaling: A comparative approach," in *Advances in the Study of Behavior*, J. S. Rosenblatt, C. Beer, M. C. Busnel, and P. J. Slater, Eds. Academic Press, 1985, vol. 15, pp. 189–244.

[23] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

[24] K. Drossos, A. Floros, K. Agkavanakis, N. A. Tatlas, and N. G. Kanellopoulos, "Emergency voice / stress-level combined recognition for intelligent house applications," in *Audio Engineering Society 132nd Convention*, 4 2012.

[25] CATT-Acoustic, *Room Acoustics Prediction and Desktop Auralization, User Manual*, CATT, Gothenburg, Sweden, 2002, v.8.