Engaging human-to-robot attention using conversational gestures and lip-synchronization

F. Cid, L.J. Manso, L.V. Calderita A. Sánchez and P. Núñez

***Abstract*—Human-Robot Interaction (HRI) is one of the most important subfields of social robotics. In several applications, text-to-speech (TTS) techniques are used by robots to provide feedback to humans. In this respect, a natural synchronization between the synthetic voice and the mouth of the robot could contribute to improve the interaction experience. This paper presents an algorithm for synchronizing Text-To-Speech systems with robotic mouths. The proposed approach estimates the appropriate aperture of the mouth based on the entropy of the synthetic audio stream provided by the TTS system. The paper also describes the cost-efficient robotic head which has been used in the experiments and introduces the use of conversational gestures for engaging Human-Robot Interaction. The system, which has been implemented in C++ and can perform in real-time, is freely available as part of the RoboComp open-source robotics framework. Finally, the paper presents the results of the opinion poll that has been conducted in order to evaluate the interaction experience.**

***Index Terms*—Robotics head, Lip Synchronization, Human Robot Interaction.**

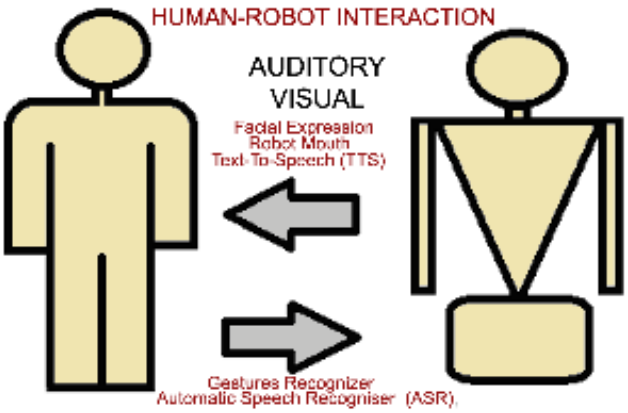

Fig. 1. HRI is usually based on visual and auditory information. For auditory information, depending on the communication direction, TTS or ASR systems are needed.

I. Introduction

DURING the last decade the robotics community interest in social robotics has grown dramatically. It is one of the fields of robotics with more practical applications. Social robots are autonomous robots that interact with humans in daily environments, following human-like social behaviors (i.e., recognizing and expressing emotions, communicating, and helping humans or other robots). During last years the use of social robots has increased for a wide variety of applications (e.g., museum guide robots[1], [2], or assistive and rehabilitation robots[3], [4]). As in other fields of application, robots can offer several key advantages for rehabilitation, such as the possibility to perform (after establishing the correct set-up) a consistent and personalized treatment without fatigue; or its capacity to use sensors to acquire data, which can provide objective quantification of recovery. However, in addition to providing physical assistance in rehabilitation, robots can also achieve personalized motivation and coaching. Thus, it is interesting to study and develop effective mechanisms of interaction between patients and robots.

This interaction between human beings and robots, usually known as Human-Robot Interaction (HRI), represents one of the biggest challenges in social robotics, resulting in new technologies and methods. Different robotic systems have been built and many studies have been conducted unveiling the importance of properly designed human-robot interaction strategies. Some of these works aim to achieve human-like robots in terms of shape [5]. Despite having a similar shape helps achieving higher empathy levels, it is not the only key factor to take into account when developing social robots. The capacity to behave similarly to human beings and to adapt to their emotional state is also a very important issue [7], [8]. Currently, different techniques are being used in order to receive input data from humans (e.g. facial expression recognition [9], skeletal modeling [10], use of corporal language [11], speech recognition [12]), but relatively little scientific research has been done regarding how robots should present the information and give feedback to their users.

In order to perceive their environment, other robots and persons, social robots are equipped with multi-modal sensors like cameras, laser range finders or microphones. Using these sensors, social robots acquire and process the necessary data for establishing communication (e.g., where the interlocutor is, or what is he/she saying or doing). On the other hand, robots working in human environments following social behaviors need methods not only to perceive but also to interact and exchange information between the source and the receiver of the message.

In order to interact with people, robots need to use different communication methods that human beings can easily receive and understand. In this regard, Natural Language (NL), in conjunction with visual information, is a very efficient method for an interaction paradigm with robots (see figure 1). Interactive NL-based communication is comfortable for humans and it is successful handling errors and uncertainties due to the fast-paced loop that such interaction provides. On the other hand, speech perception involves information from more than one

F. Cid, L.J. Manso, L.V. Calderita, A. Sánchez and P. Núñez are with University of Extremadura.
E-mail: fecidb@alumnos.unex.es

sensory modality. In particular, visual information has been proven to be strongly linked with hearing when recognizing speech (McGurk effect [13]). Therefore, it is very likely that mouth synchronization will also help in order to keep the attention of the users in what the robot says. The hypothesis of this proposal is that the HRI experience can be improved using the visual information provided by a robotic mouth whose movements are synchronized with the synthetic voice. Besides, it is analyzed the use of conversational gestures for engaging human-robot interaction.

The perception of the state of the robot and the understanding the voice messages it synthesizes can be improved providing additional information. There are two common information sources for this: **a)** auditory cues (e.g., pitch, pauses, emphasis); and **b)** visual gestures (e.g., lip movements, facial expressions, neck movements). This information allows message senders to acknowledge their emotional state, their intentions and even to transmit concepts. Thus, it is very important to accompany voice with visual feedback and auditory cues in order to ease the correct interpretation of the message to transmit.

This paper presents a robotic head and a synchronization algorithm on a robotic mouth that can perform in real-time with different TTS systems. This algorithm is based on a synchronization algorithm that uses the entropy of the synthetic audio stream for estimating the level of aperture of the robotic mouth. The robotic head, which has a very cost-efficient design, has been included in Ursus social robot, a therapy robot with the shape of a teddy bear [6]. Ursus was designed to improve the therapy of children with developmental disorders like cerebral palsy by making a game of the therapy. Achieving an entertaining therapy for children helps them keeping their attention, which improves the results.

In order to evaluate the initial hypothesis, an opinion poll was conducted with different participants, both roboticists and non-roboticists. The poll took into account: **1)** the impact of the different mouths used in the poll, physical and simulated ones; **2)** how the different TTS systems for voice synthesis influenced user experience; **3)** the impact of the different synchronization algorithms described in the literature [14]; and **4)** how the body language improves the communication of concepts and emotions to the human being. Other factors such as the level of engaging, understanding or acceptance were also evaluated.

The rest of this paper is organized as follows. Section II introduces the state-of-art of the different HRI techniques and their evolution. Section III presents an overview of the proposed system. Next, the robotic mouth designed is described in section IV. Section V presents the synchronization algorithm, describing in detail the different stages of the process. Finally, the results of the experiments proposed in this paper and the conclusions are detailed in section VI and VII, respectively.

II. Related work

Affective communication has been the core topic of different social robotics works. It aims to reduce the communication gap between humans and robots not just by using natural language but also by providing robots with human-like gestures and, to some extent, shape. These techniques allow roboticists to achieve stronger human-robot empathy [15]. Moreover, it is easier for humans to interact with agents with similar characteristics (e.g., appearance, communication mechanisms, gestures). The use of speech-guided dialogue to teach robots [16] allows roboticists and end-users to control and interact with robots using natural language. The first step to achieve this kind of interaction is to be able to send and receive messages through a media that humans can understand. This is done by using technologies such as audio synthesizers (TTS) [17] and speech recognition systems (ASR) [18]. In the last years, these systems are becoming very common in social robotics [19], [20], [21], [22].

Lip synchronization in robotics looks for matching lip movements with the audio generated by the robot. The use of different lip synchronization algorithms not only are limited to use in robotics, but also to the lip animation in virtual models used in HRI systems with computers. These systems allow the user to interact with virtual models through speech and some cases through body language [27], [28], [29]. Several works use synchronization algorithms based directly on the use of audio phonemes to determine the levels of mouth aperture [23], [24]. These approaches require additional information such as dictionaries of phonemes. In this paper it is presented a synchronization algorithm based on the entropy of the synthetic audio signal provided by the TTS system. Despite being a different problem, some authors have successfully used entropy for automatic speech recognition [30], [31], [32] (especially in noisy environments).

Finally, TTS-enabled robots can provide information to humans, but they usually are unembodied monotonous voices, lacking of emotion, pitch changes and emphasys. In [38] it is presented a study of how the prosody of speech influences auditory verbal communication. Similar to [25] and [26], the proposed approach evaluates the different aspects of speech-based interaction.

III. Overview of the System

The main goal of the proposed system is the design of a robotic mouth and the control of a robotic head in order to provide visual information for helping the understanding of the messages synthesized by robots. The mouth is governed by a TTS-lip synchronization algorithm. Thus, the lips move according to the synthesized voice generated using a Text-to-Speech system. This setup helps keeping the attention of social robot users. Figure 2 illustrates an overview of the system. As it is shown in the figure, it is constituted by two layers, hardware and software. The hardware is composed of a previously made robotic head with three degrees-of-freedom, a speaker in order to hear the voice of the robot and the proposed mouth.

IV. Robotic head

The robotic head used in this paper consists of two elements: a neck and a robotic mouth. The robotic neck is driven by three Dynamixel RX-10 servos, allowing pitch, roll and yaw

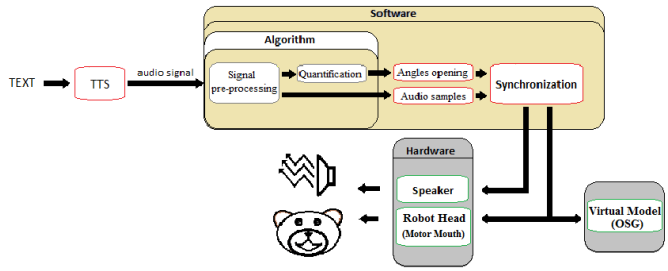

Fig. 2. Overview of the proposed system in this paper. Both, software and hardware layer are depicted in the figure.

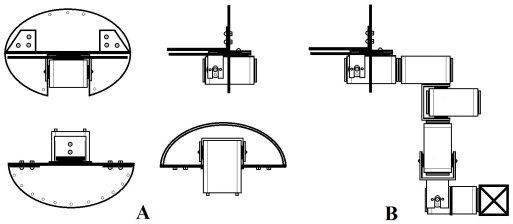

Fig. 3. Different views of the mechanical system. From left to right and from top to down: frontal, profile, top and bottom views.

Fig. 4. On the left hand side is shown, the robotic mouth mounted on Ursus2. On the right hand side is illustrated, the mouth system separated from the rest of the robot, as it is described in section VI

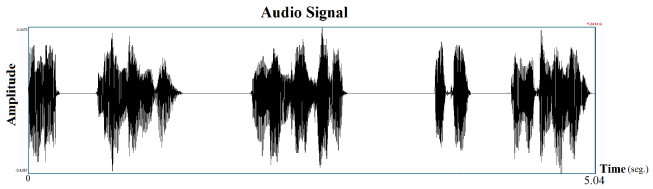

Fig. 5. Audio signal waveform.

movements. The key design considerations for the robotic mouth, which was specially built for this work, are: i) the efficiency of the mechanical system, considering a reasonable range of aperture of the mouth; ii) the suitability of the mouth for its use on the Ursus therapy robot and; iii) the overall price of the mouth. The CAD design of the robotic head (including neck and mouth) is illustrated in figure 3.B. The mechanical structure of the robotic mouth consists of three aluminum planar pieces, corresponding to the chassis of the mouth (figure 3.A), upper and lower lips and a Dynamixel RX-10 servo. The upper aluminum piece is fixed, while the lower lip is moved by the motor. The mouth aperture was set to range between 0 and 45 degrees. Finally, the mechanical pieces are covered by a fabric similar to those used in teddy bears (figure 4).

V. Synchronization Algorithm

A. Text To Speech System

Usually, speech synthesis systems are used in order to directly take the audio output to the speakers. In our case, since we want to make sure that the audio output is synchronized with the mouth movements, the TTS system does not have access to the speakers and it only generates the output audio file. These audio files are then concurrently used for producing both the mouth movements and the audio output.

The proposed lip synchronization algorithm is independent of the Text-to-Speech system. In the experiments shown in this section Verbio TTS system has been used [17] in order to illustrate partial results. This system can generate audio output for different languages, using various audio formats such as OGG or WAV, and allowing adaptive and dynamic intonation.

In particular, the following setup has been used in the approach:

- Language Spanish and English.
- File format OGG.
- Sample rate $Fs = 16Khz$.

Figure 5 illustrates the audio signal obtained for the text: "Hello, my name is Ursus, tell me what is your name".

In section VI, different TTS systems (proprietary and free/libre software) are used for comparison purposes (Verbio, Festival, Ivona and Acapela) in order to demonstrate the correct operation of the lip synchronization algorithm for each one of them. Independent of the TTS system, the proposed algorithm only needs: the sample rate and the output audio file.

B. Signal preprocessing

As it was introduced, mouth movements are based on the entropy level of the audio signal, whose value is calculated on-line for every time window. The input of the algorithm is the audio signal $X(t)$, which has a length of $Fs \cdot T$, where Fs is the sample rate and T the duration of the whole audio signal.

$$X(t) = [0, ..., F_s \cdot T - 1] \quad (1)$$

and obtain the entropy of the windows, the following steps must be taken previously:

- **A)** Obtain the absolute value of the audio signal:

$$V(i) = |X(t)| \quad (2)$$

- **B)** Windowize the signal vector, since the entropy is computed for each window separately.

Time windows have a length of a tenth of a second. It is an adequate length given the nature of the signal (i.e., phonems are usually about a tenth of a second long) and the response time of the motors.. The signal preprocessing step is shown in figure 6.

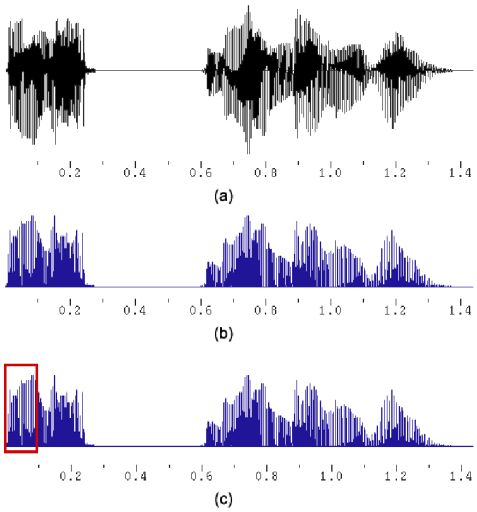

Fig. 6. Signal preprocessing: a) initial audio signal, b) absolute value of the audio signal, b) example of a time window.

C. Quantification

In this work an entropy-based algorithm is proposed in order to set the mouth aperture of the robot given the current audio stream. Since the audio stream is synthetic, it can be safely assumed that the audio is noise free. Thus, the algorithm provides a mouth aperture proportional to the audio entropy for each of the time windows.

Entropy quantifies the existent amount of information in a given signal. Given a set of different samples $1...n$ of a random variable X (which can be interpreted as a signal), the amount of information on it (measured in bits) can be computed as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i) \tag{3}$$

where x_i is the nth measurement and $P(x_i)$ the probability of finding that measurement within the time window.

Finally, the angle sent to the motor is proportional to the entropy level:

$$angle \propto entropy$$

In our experiments, the proportional constant was experimentally set to 1.5. It might vary depending on the text-to-speech system.

D. Synchronization

The opening levels computed by the algorithm must be synchronized with the audio sent to the speakers. This synchronization is made using the same audio libraries which are used for playback, processing and quantification the audio signals. Thus, the audio samples are simultaneously processed by the audio library and the angles calculated in each time windows are sent to the motors of the robot mouth (see figure 2). Thus, communication delays between the computer and the motors are reduced and the synchronization results are improved.

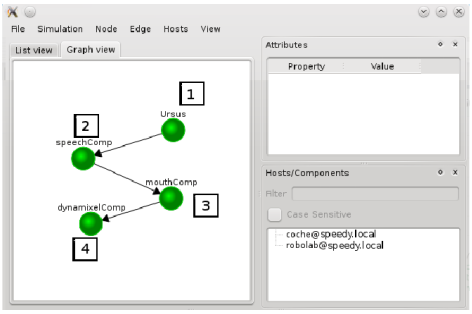

Fig. 7. Screenshot of the *RCManager RoboComp* tool. Ursus main component (1), which is connected to the TTS system (2). The component which transforms the sound in motor movements is labeled as 3. Component moving the servomotor (4).

E. RoboComp Components

The software to control our system is built on top of the robotics framework *RoboComp* [35]. Making use of the components and tools it provides and its communication middleware we developed an easy to understand and efficient architecture (see figure 7).

The main component of the proposed system is ursusComp. It is connected, directly or indirectly, to the rest of the software components controlling Ursus: camera, robotic arms, tracker, etc (figure 7). Not all components have been included in the diagram in order to make it simple. The sentences that Ursus tells its patients to encourage them during their therapy are sent to speechComp (see figure 7-(2)). Then speechComp transforms the sentences into sound using the specific TTS system (e.g. Festival, Verbio). After that, mouthComp (figure 7-(3)) receives the sound and send the motor commands using the synchronization algorithm. Finally, the motor commands are received and executed by dynamixelComp.

Since the system was designed an implemented using component oriented design/programming, these components can be easily used for other purposes, which is a very important feature in robotics development.

VI. Experimental Results

One of the main goals behind the development of the robot head and the synchronization algorithm is to use them as an improvement for Human Robot Interaction. The initial hypothesis was that the use of a robotic mouth moving according to the synthetic voice generated by a Text-To-Speech system allows a) robots to maintain the attention of their users while talking and, b) human beings to interact more efficiently with robots. The idea is that robots equipped with motorized mouths can achieve a better interaction than those which are not. The second hypothesis is that the use of head body language is useful in order to successfully transmit concepts or emotions.

There exist different approaches to evaluate the performance of social robots when interacting with humans. In addition to evaluating the synchronization algorithm, it is also interesting and necessary to analyze how the proposed robot mouth affects humans. For this purpose, different works and researchers propose the use of quantitative measures of human attention or body movement interaction between robots and humans [34],

Fig. 8. First version of the therapist robot Ursus.

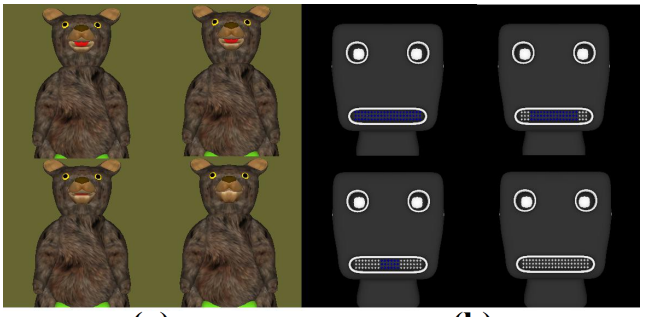

Fig. 9. Screenshots of different mouths. a) Virtual Robot Ursus b) Robotic Model based on LED matrix mouth. Both models are shown in four different positions.

[33]. In this paper, acceptance, engaging and understanding are three factors to be measured in the HRI context. These factors are evaluated using pool-based methods, where the opinion of the user is surveyed.

Thus, the performance of the proposal has been evaluated based on the impression of the participants regarding the synchronization algorithm and the robotic head according to: 1) the difference in perception between a physical robotic mouth and a simulated one; 2) how the different TTS systems for voice synthesis influenced user experience; 3) the impact of the different synchronization algorithms described in the literature [14]; and 4) how the body language improves the communication of concepts and emotions to humans.

A. Robot platform Ursus

Ursus is an assistive robot developed in the Laboratory of Robotics and Artificial Vision of the University of Extremadura (Cáceres, Spain) in collaboration with the Virgen del Rocío Hospital (Sevilla, Spain). It is designed to propose games to children with cerebral palsy in order to improve their therapy. It will also be used as a tool for their therapists to adjust therapy to the needs of the different patients. In order to make it visually pleasant for children, it has a friendly height and has been wrapped into the covering tissue of a teddy bear.

Patients can get feedback of the status of the exercise in real-time by looking at an image that the robot projects on the wall. Along with the messages the robot sends to the patients, this information encourages children to improve the execution of the exercises. Figure 8 illustrates the first version of Ursus. Ursus is composed of two arms, both of four degrees of freedom (DOF), mounted on a fixed torso. These are used so that patients can see how the robot perform the exercise and try to reproduce the movement. A regular USB camera is located in the neck of the robot to capture the arm movements of the users, allowing the robot to provide appropriate feedback about their performance. The speaker and the computer are located on the base of the robot.

B. Comparative study

Social robotics enables robots to interact with diverse groups of people, through simple and friendly means of communication such as as speech [23] [20]. A comparative survey was conducted to assess various aspects of the mouth through a series of questions with a response on a linear scale of 1-5. For the study has been selected 15 persons, each one with different degrees of knowledge about robotics. These degrees of knowledge can be divided into three categories: high (5 persons), medium or moderate (4 persons) and low (6 persons).

The evaluated items were divided into four groups: robotic mouths, TTS softwares, lip synchronization algorithms and body language, which were compared in order to determine the best in each group.

The following elements were evaluated by the participants:

- **A)** Natural behavior
- **B)** Expressiveness
- **C)** Attention engaging capacity
- **D)** Message understanding

Different questions were used in order to obtain an average score for each study as:

- Does the mouth seem to move naturally?
- Does the mouth seem expressive?
- Did the mouth capture your attention?
- Did the mouth, directly or indirectly, help you to understand the message?

Were repeated using different sentences in order to obtain an average value.

C. Comparative study of different robot mouths

The robotic mouth was compared with two different designs used for researching in Human Robot Interaction. First, Figure 9.a illustrates the robot Ursus virtual model created in 3D Studio Max with a mouth, which it is moved according to the proposed synchronization algorithm. The second robotic mouth included in this comparative study is based on other research works that use a LED matrix [14] (see figure 9.b). Instead of developing the hardware LED matrix it was also simulated. It consists on a 21x3 matrix whose elements are enabled according to the synchronization algorithm. Thus, they are turned on as entropy increases. Both mouths are displayed on the screen monitor using a size similar to the Ursus one. In figure 4 it is shown the set up used for evaluating the robotic mouth.

Mouth	Questions			
	A	B	C	D
Animated mouth	67%	68%	69%	78%
Led mouth	42%	46%	59%	73%
Robotic mouth	74%	66%	74%	64%

TABLE I
COMPARISON OF THE DIFFERENT MOUTHS TESTED

TTS	Questions			
	A	B	C	D
Verbio	52%	46%	52%	72%
Festival	60%	56%	52%	80%
Acapela	68%	72%	68%	72%
Ivona	64%	60%	56%	68%

TABLE II
COMPARISON OF THE DIFFERENT TTS

TTS/Robotic mouth	Questions			
	A	B	C	D
Verbio	74%	66%	74%	64%
Festival	50%	45%	50%	60%
Acapela	80%	75%	80%	70%
Ivona	65%	60%	70%	60%

TABLE III
COMPARISON OF THE DIFFERENT TTS IN THE ALGORITHMS OF SYNCHRONIZATION

Algorithms synchronization	Questions			
	A	B	C	D
Entropy	80%	80%	80%	64%
Random	40%	44%	40%	36%
Binary	48%	44%	48%	48%

TABLE IV
COMPARISON OF THE DIFFERENT TTS IN THE ALGORITHM OF SYNCHRONIZATION

The same pool and participants were used for evaluating the features of the different robotic mouths. Results are summarized in table I.

As shown in table I, the robotic mouth presented in this paper performs better compared with other mouths in elements such as: natural Behavior and Attention Engaging capacity, resulting in a greater interaction and attention by the survey participants. In addition to the robotic mouth, the survey shows that the animated mouth presents excellent results in items such as: expressiveness and response in the Message Understanding.

D. Comparative study of the different Text-To-Speech systems

This section describes the evaluation of different TTS systems. In this study four different TTS systems have been used: Verbio, by Verbio Technologies; Festival, by the University of Edinburg; Acapela, by the Group Acapela; and Ivona, by the company Ivona Software.

One of the main aspects to take into account when using a TTS system is the output sample rate. In this study, for each TTS, the following sample rates were used.

- Verbio: $16Khz$
- Festival: $44Khz$
- Ivona: $22Khz$
- Acapela: $22Khz$

The algorithm can be used with any TTS system, as long as it complies with certain parameters such as audio sampling frequency or the ability to produce output files.

For the evaluation of the TTS systems, the questions specified in section VI-B have been used. The evaluation results of the TTS systems are summed up in table II.

The results shown in the table II, demonstrate that the Acapela TTS software performs better than other TTS in aspects such as naturalness or expressiveness. In addition to evaluating the synthesizer, the poll took into consideration the performance of each TTS with the robotic mouth and the algorithm of synchronization, that is the key of this work. The corresponding results to the use of the TTS with the robotic mouth are shown in table III.

Table III shows that the best achieved performance is produced by Acapela in conjunction with the proposed synchronization algorithm and the robotic mouth developed for this paper. Demonstrating how the synchronization algorithm helps improve speech perception, and allowing the user to be able to perceive of simplest form, the elements evaluate in this survey, such as attention engaging capacity, expressiveness or the natural behavior.

E. Comparative study of different synchronization Algorithms

Finally, the comparative study allowed to evaluate the synchronization algorithm compared to other algorithms, such a binary pulse delivery aperture (mouth opened if there is sound) and other that controls movement through random levels of mouth aperture.

For the evaluation of these synchronization algorithms a survey was made with the questions of the subsection VI-B. Results have been summarized in table IV.

The results of the survey demonstrate in the table IV, that the synchronization algorithm based on entropy provides a better user experience in comparison to similar algorithm in all the elements evaluated by survey. Besides, being the best performing in speech perception, it has other features that make it useful for social robotics, as its ability to work with any TTS software.

F. Comparative study of the used of the body language

This comparative study evaluated the impact of body language in human-robot communication. An experiment was conducted with voice messages from a TTS system with modified parameters, as emphasis, pitch, pauses, among others. These changes affected the entire sentence, at the same time that a few words. The set of parameters of the algorithm was not modified, except the size of the time windows after modifying the speed of the TTS. In addition, a second experiment that used not only the voice message, but also a series of movements (body language) to express a concept as doubt, a question and anger. Both two experiments have been conducted with a virtual robot (specifically the robot Ursus).

The results of the survey carried for the two experiments can be seen in Table V:

Body language	Questions			
	A	B	C	D
Only voice	54%	60%	63%	75%
Voice and Move	56%	65%	76%	75%

TABLE V
COMPARISON OF THE USED OF BODY LANGUAGE.

In table V, the experimental results have shown how body language, in comparison to the dialogue based only in the speech, provides a better performance in elements as: natural behavior, expressiveness and attention engaging capacity. Thus, the importance of body language in the interaction of robots with users is demonstrate, giving rise to a natural language based on the motion and speech by the robots from humans, but allowing similar communicate.

VII. CONCLUSION

Social robots need to communicate properly in order to improve their interaction skills with people. In this respect, both visual and auditory sources must be taken into account as it was demonstrated by McGurck[13]. The use of visual feedback (i.e., head and mouth movements) can be used, not only to improve understanding, but also to achieve higher levels of attention and empathy.

The results provided by the survey demonstrate that the proposed algorithm has several advantages over the state-of-the-art algorithms. Moreover, our algorithm performs in real-time and does not require additional training such as other approaches [23], [24].

Currently, the RoboLab group is working in order to be able not only to provide speech information to the user but also to receive it. By removing the need to make the user move (i.e. in order to touch a touchscreen), is expected that user experience will be dramatically enhanced.

ACKNOWLEDGMENT

This work has been partially supported by the Junta de Extremadura project IB10062 and by the Spanish Ministry of Science and Innovation with grant IPT-430000-2010-002.

REFERENCES

[1] W. Burgard, A. B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer. “Experiences with an interactive museum tour-guide robot”.In *Artificial Intelligence (AI)*, pp. 3-55. 2000.

[2] F. Faber, M. Bennewitz, C. Eppner, A. Gorog, “The humanoid museum tour guide Robotinho”. In *Proc. IEEE International Symposium on Robot and Human Interactive Communication* (RO-MAN). September 2009.

[3] X. Ma and F. Quek, “Development of a Child-Oriented Social Robot for Safe and Interactive Physical Interaction”. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* Taiwan, pp. 2163-2168, October 2010.

[4] T. Mukai, S. Hirano, H. Nakashima, Y. Kato, Y. Sakaida, S. Guo, and S. Hosoe, “Development of a Nursing-Care Assistant Robot RIBA That Can Lift a Human in Its Arms”, In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taiwan, pp. 5996-6001, October 2010.

[5] P.Breen, E. Bowers, W. Welsh, “An investigation into the generation of mouth shapes for a talking head”. In *Proc. of ICSLP 96*, USA, pp. 2159-2162, October 1996.

[6] L.V. Calderita, P. Bachiller, J.P. Bandera, P. Bustos, and P. Nunez, ”MIMIC: A Human motion imitation component for RoboComp”. In *Proc of Int Workshop on Recognition and Action for Scene understanding.*. 2011.

[7] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, “Development of the Face Robot SAYA for Rich Facial Expressions”. In *Proc. of 2006 SICE-ICASE International Joint Conference* Korea, pp. 5423-5428, October 2006.

[8] M.J. Matarić , J. Eriksson , D.J. Feil-Seifer1 and C.J. Winstein. “Socially assistive robotics for post-stroke rehabilitation”. In *Journal of NeuroEngineering and Rehabilitation*. 4:5. 2007.

[9] J.A. Prado, C. SimplÃÑcio, N.F. Lori and J. Dias, ”Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction”, In *International Journal of Social Robotics*, vol. 4, no. 1, pp. 29-51, December 2011.

[10] J.P. Bandera, ”Vision-Based Gesture Recognition in a Robot Learning by Imitation Framework”, Ph.D. Thesis, University of Malaga, 2009.

[11] A. Aly and A. Tapus, ”Speech to Head Gesture Mapping in Multimodal Human-Robot Interaction”, In *Proc. of the 5th European Conference on Mobile Robots ECMR 2011* Sweden, pp. 101-108, September 2011.

[12] C. Breazeal and L. Aryananda, ”Recognition of Affective Communicative Intent in Robot-Directed Speech”, Artificial Intelligence, pp. 83-104, 2002.

[13] T. Chen, “Audio-Visual Integration in multimodal Communication”. In *IEEE Proceedings*, May, 1998.

[14] M.K Lee, J. Forlizzi, P.E. Rybski, F. Crabbe, W. Chung, J. Finkle, E. Glaser, S. Kiesler, “The Snackbot: Documenting the Design of a Robot for Long-term Human-Robot Interaction”. In *Proc. of HRI 2009*, pp. 7-14, 2009.

[15] R. W. Picard, “Affective Computing”. MIT Press, pp. 88-91, 2000.

[16] J. Gómez, A. Ceballos, F. Prieto, T. Redarce, “Mouth Gesture and Voice Command Based Robot Command Interface”. In *Proc. of IEEE International Conference on Robotics and Automation*, Japan, pp. 333-338, May. 2009.

[17] Verbio Technologies, “Text to Speech (TTS) and Speech Recognition (ASR)”. Available at: *http://www.verbio.com*.

[18] S. Anderson, D. Kewley-Port, “Evaluation of Speech Recognizers for Speech Training: Applications”, In *IEEE Transactions on speech and audio processing*, VOL. 3, NO. 4, pp. 229-241, July 1995.

[19] C. Jayawardena, I. H. Kuo, U. Unger, A. Igic, R. Wong, C. I. Watson, R. Q. Stafford, E. Broadbent, P. Tiwari, J. Warren, J. Sohn and B. A. MacDonald. “Deployment of a Service Robot to Help Older People”. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems* Taiwan, pp. 5990-5995, October 2010.

[20] C. Shi, T. Kanda, M. Shimada, F. Yamaoka, H. Ishiguro and N. Hagita “Easy Development of Communication Behaviors in Social Robots”. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* Taiwan, pp. 5302-5309, October 2010.

[21] W. Zhiliang, L. Yaofeng, J. Xiao, “The research of the humanoid robot with facial expressions for emotional interaction”. In Proc. *First International Conference on Intelligent Networks and Intelligent Systems*, pp. 416-420. 2008.

[22] P. Rybski, K. Yoon, J. Stolarz, M. Veloso, “Interactive Robot Task Training through Dialog and Demonstration”. In *In Proc. of HRI 2007*, USA, 2007.

[23] K.-geune Oh, C.-yul Jung, Y.-gyu Lee, and S.-jong Kim, “Real Time Lip Synchronization between Text to Speech(TTS) System and Robot Mouth”. In *Proc. of IEEE International Symposium on Robot and Human Interactive Communication*,Italy, pp. 620-625, September. 2010.

[24] F. Hara, K. Endou, and S. Shirata, “Lip-Configuration control of a Mouth robot for japanese vowels”. In *Proc. IEEE International Workshop on Robot and Human Communication* pp. 412-418, 1997.

[25] A. Austermann, S. Yamada, K. Funakoshi, M. Nakano, “Similarities and Differences in Users Interaction with a Humanoid and a Pet Robot”. In *Proc. 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* pp.73-74, 2010.

[26] H. Kamide, Y Mae, T. Takubo, K. Ohara and T Arai. ”Development of a Scale of Perception to Humanoid Robots: PERNOD”. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* Taiwan, pp. 5830-5835, October 2010.

[27] S. DiPaola, A. Arya, J. Chan, ”Simulating Face to Face Collaboration for Interactive Learning Systems”, In *In Proc: E-Learn 2005*, Vancouver, 2005.

[28] K. Waters and T.M. Levergood, ”DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces”. In *MULTIMEDIA TOOLS AND APPLICATIONS* Vol. 1, Number 4, pp. 349-366, 1995.

[29] Kaihui Mu, Jianhua Tao, J. che and M. Yang, "Real-Time Speech-Driven Lip Synchronization", In *4th International Universal Communication Symposium (IUCS), 2010* , Beijing ,pp. 378-382, 2010

[30] J-L. Shen, J-W. Hung, L-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments", In *ICSLP-1998*, 1998.

[31] J.C. Junqua, B. Mak, and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", In *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No.3, pp. 406-412, Apr. 1994.

[32] M.H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech", In *Speech Communication*, Vol. 8, pp. 45-60, 1989.

[33] A. Tapus and M.j. Mataric, "Emulating Empathy in Socially Assistive Robotics", In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*, Stanford, USA, March 2007

[34] A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll and L. Hall, "Caring for Agents and Agents that Care: Building Empathic Relations with Synthetic Agents", In *Third International Joint Conference on Autonomous Agents and Multiagents Systems*, Vol. 1, pp. 194-201, New York, USA, 2004.

[35] L.J. Manso, P. Bachiller, P. Bustos, P. Nuñez, R. Cintas and L. Calderita. "RoboComp: a Tool-based Robotics Framework". In *Simulation, Modeling and Programming for Autonomous Robots* (SIMPAR). Pages 251-262. 2010.

[36] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive Robotics: The influence of robot gender on human behavior". In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2563-2568, October 2009.

[37] H. Song, and D. Kwon, "Design of a Robot Head with Arm-type Antennae for Emotional Expression". In *International Conference on Control,Automation and System. ICCAS '07*, pp. 1842 - 1846, October 2007.

[38] J. Cahn. "Generation Expression in synthesized speech". Master's Thesis, MIT Media Lab. 1990.