Journal of Intelligent & Robotic Systems manuscript No.
(will be inserted by the editor)

Part-of-Speech and Prosody-based Approaches for Robot Speech and Gesture Synchronization

L.Pérez-Mayos · M.Farrús · J.Adell

Received: date / Accepted: date

Humanoid robots are already among us and they are beginning to assume more social and personal roles, like guiding and assisting people. Thus, they should interact in a human-friendly manner, using not only verbal cues but also synchronized non-verbal and para-verbal cues. However, available robots are not able to communicate in this multimodal way, being just able to perform predefined gesture sequences, hand-crafted to accompany specific utterances. In the current paper, we propose a model based on three different approaches to extend humanoid robots communication behaviour with upper body gestures synchronized with the speech for novel utterances, exploiting part-of-speech grammatical information, prosody cues, and a combination of both. User studies confirm that our methods are able to produce natural, appropriate and good timed gesture sequences synchronized with speech, using both beat and emblematic gestures.

1 Introduction

Humanoid robots are already beginning to function in shared spaces with humans, assuming roles that are more social and personal. Robots that guide, assist and help people should communicate in a human-friendly way, using not just speech but also gestures to improve the communication process [2]. Verbal and textual information is not enough to characterize human communication: when we speak, we use facial gestures to express emotions and hand gestures to better describe and shape the concepts we are talking about, increasing communication efficiency [8]. Those gestures are called co-speech or co-verbal gestures, and they have been shown to have positive effects on listener behaviour, helping the listener to pay attention longer and recall more information [21,5,8]. Moreover, research has shown that multimodal communication, e.g. using simultaneously voice and gestures to transmit information, benefits both speakers and listeners [30].

L. Pérez-Mayos
Department of Information and Communication Technologies, UPF, Barcelona
E-mail: laura.perezm@upf.edu

M. Farrús
Department of Information and Communication Technologies, UPF, Barcelona
E-mail: mireia.farrus@upf.edu

J. Adell
Verbio Technologies S.L., Barcelona
E-mail: jadell@verbio.com

Even though humanoid robots development has moved forward dramatically in the last few years, they are not yet able to perform synchronized gestures along with their speech automatically: they either perform random gestures or perform predefined gesture sequences, handcrafted to accompany specific utterances [22]. Thus, they are not able to produce gestures to accompany novel utterances, resulting in a poor communication experience for the people who interact with them. The complex task of designing robot gestures builds upon the human gesture characterization. In the early nineties, [21] proposed a gesture classification according to the semantic functions that has been widely used ever since. According to such classification, *emblematic*, *metaphoric*, *deictic* and *iconic* gestures relate closely to semantics –the study of the meaning–, while *beat* gestures are closely related to morphology, –which studies the structure of the morphemes contained in the language and other linguistic units. Beat gestures do not directly convey meaning, but they refer to the speaking process by synchronizing with prosodic events in speech. Beat gestures have been found to contribute to the perceived prominence of temporally aligned speech, and can function in the sense of audiovisual prosody [30].

For the co-speech gesture synchronization, [21] suggests the following synchronization rules:

1. The *phonological* rule predicts that a gesture stroke should occur before the most prominent syllable.
2. The *semantic* rule predicts that co-occurring gestures and speech relate to the same idea unit.
3. The *pragmatic* rule predicts that co-occurring gestures and speech have the same pragmatic function.

A human gesture is decomposed into three phases of motion, namely preparation, stroke, and retraction [17]. The stroke is the fragment of higher gestural effort, whilst the preparation and retraction elements consist of moving the arms to and from the rest position, to and from the start and end of the stroke. Moreover, gestures have a frame of reference called the gesture space [21], which organizes the space in front of the speaker's body into positions, regions and directions. Most gestures are performed in central gestural space, a sphere-shaped area in front of the speaker's upper part of the body, below the neck and between the shoulders and elbows. Gestures occurring outside this area are said to be produced in peripheral gestural space.

Several works have dealt with the use of prosody to generate and synchronize co-speech gestures [16,2, 1,3,4]. Other works have also used the grammatical part-of-speech information, but only to provide timing information of spoken works in a text-to-speech system [25]. In this work, we go beyond the existing works and we propose a model that uses prosody information for robot gesture generation and synchronization, as well as the part-of-speech information and the combination of both. The main goal is to extend humanoid robots communicative behaviour with upper-body gestures automatically synchronized with the speech. We design a method to characterize, design and store human gestures so they can be performed by REEM-C, a full-size biped humanoid robotics research platform [1] developed by PAL Robotics [2]. The three proposed approaches in our model aim at synchronizing gestures with the speech, and at exploring whether a subset of gesture types can be representative enough to reproduce a natural human-like communication behaviour in a robot.

The remainder of this paper is organized as follows. In section 2, different gesture generation and co-speech generation methods are overviewed, focusing on those using prosodic and linguistic information. Section 3 presents the experimental setup under which this research was conducted, including the technical framework for gesture generaiont. In section 4 we present the three different approaches for co-speech gesture synchronization addressed in this work, based on: (a) part-of-speech grammatical information, (b) prosody, and (c) a combination of both. In section 5, the results of the different approaches are presented and discussed. Finally, in section 6 the conclusions are drawn along with the directions for future research.

[1] `http://pal-robotics.com/es/products/reem-c`

[2] `http://pal-robotics.com`

2 Co-speech Gesture Generation and Synchronization

This section overviews the related work on human gesture communication, and on the use of prosodic and linguistic information in gesture generation and synchronization of virtual agents and robots.

2.1 Human gesture communication

It is well-known that visual information is a relevant aspect to achieve a fully speech perception, and such relevant visual information includes not only lip movements, but also face, head, and body gestures [24, 18]. Besides, human-robot interaction and communication can be established through verbal, non-verbal, and para-verbal cues [2]. Para-verbal and non-verbal communications are naturally synchronized, linked in both production and perception [30,1,13]. In this light, it is also well-known that prosody plays an important role in daily-life communication [26,31], being constantly adjusted to the particularities of the speaker. Such prosodic variations are manifested through the three prosody elements: intonation, rhythm, and stress, having a special role in speech and gestures alignment [30]. Audiovisual prosody, thus, becomes an essential aspect to model human communication and human-computer interaction [18].

Humans use different types of gestures to communicate with each other. Moreover, those gestures accompanying speech have the potential to display thoughts that are not conveyed in speech [15]. In the early nineties, [21] proposed a gesture classification according to their semantic functions that has been widely used in research: (a) *emblematic* gestures —those that bear conventionalized meaning, e.g., "thumbs up"—, (b) *beat* gestures —simple fast hand movements, with no concrete meaning, synchronized with prosodic events in speech—, (c) *metaphoric* gestures — those that resemble abstract content and ideas being explained—, (d) *deictic* gestures —those that point out location in space, this being either conceptual of concrete—, and (e) *iconic* gestures —those that resemble a certain physical aspect of the conveyed information; e.g., those used to describe the shape of an object or the direction of a movement.

2.2 Prosody and linguistic information for virtual agent gesture generation and synchronization

Experimenting with humanoid robots is hard and expensive, since robots are usually characterized by a low degree of expressiveness. Therefore, most of the research on speech–gesture synchronization, and audiovisual prosody is conducted using virtual agents. This specially affects the research focused on facial movements, which mostly relies on eyes and lips [32,11]. In fact, the majority of gestures are performed in a sphere-shaped are in front of the speaker's upper part of the body, called the central gestural space [30]. For this reason, all the current research regarding co-speech gesture generation is focused on the upper part of the body.

Generating co-speech gestures for virtual agents is composed of several tasks: (1) deciding which gestures will the robot be able to perform, (2) encoding and storing those gestures in a database, and (3) synchronizing them with the speech. These tasks can be approached in several ways. In *tele-operated* approaches, for instance, a tele-operated robot performer is used to show that speech emphasis has a similar impact on speech understanding from a human or robot performer, but beat gesture emphasis has significantly less effect when performed by a robot than when performed by a human, being gestures that accompany emphasis and timing of talk and relevant for the perception of robot performance naturalness [7,6]. *Motions automatically extracted from video input* is another approach that used a fully parameterized Hidden Markov Model (HMM) to drive automatically virtual agent behaviours from speech signals, extracting subsets of facial motion features corresponding with the Facial Animation Parameters [11,27]. In other works, raw audio and video training inputs are analyzed to extract relevant characteristics (e.g., pitch-intensity curves for voice and motion curves for gesture), and then the extracted prosody patterns and

the motion curves are aligned separately [2,4]. Alternatively, a corpus of motion capture and audio data can be processed to build a probabilistic model that correlates gesture and prosody [20]. In the *Database Annotation* approach, a database is annotated with gestures, constructing a gesture dictionary, together with a set of gestural dimensions such as amplitude, fluidity, power and speed, allowing to increase the expressiveness of the gestures [32,23,19]. Finally, *Expert Systems* are used when no corpus is available, by recording an ad hoc corpus to obtain animations and their corresponding video recordings, together with the level of intensity and strength of gestures, lexemes and gesture phases denoting the start, main stroke and retraction of gestures over time, or with different bags of words to enable gesture primitives selection [12,29,25].

Once the gestures are determined, a gesture selection must be performed and synchronized together with the corresponding speech. To this end, the use of HMM and prosodic information is a very convenient approach when a multimodal corpus is available [11,4,2,20]. Alternatively, hybrid approaches can be used by combining HMM models with rules and global statistics to construct a mapping from speech prosodic information to facial gestures [32]. Also, machine learning techniques can be used along with prosodic information to generate co-speech hand gestures for embodied agents [10], or using a text-to-speech (TTS)-driven non-verbal behaviour system for co-speech gesture synthesis based on several linguistic features that are extrapolated from arbitrary input text sequences and prosodic features such as pitch, stress, emphasis, etc. [23]. Alternatively, learning prosody-based motion generators for virtual agents with recorded speech can be used to build a model that learns the temporal relation of human gesture and the relation between prosody and motion dynamics [9]. Finally, *expert systems* can also be used for co-speech gesture synchronization [12,29,25].

2.3 Prosody and linguistic information for robot gesture generation and synchronization

The techniques used in the virtual agent domain are usually applicable to humanoid robots. Actually, some works use platforms aimed at controlling virtual agent systems for robots [14]. However, the lack of visible facial expression underscores the need for more expressive bodily communication, using not just the arms and hands but also the torso and head orientation [25]. Enhancing humanoid robots communicative behaviour with co-speech gestures helps users pay attention longer and recall more information [21,5,8]. In both assistance and social roles gesturing as naturally as possible helps to promote the users engagement and keep their attention longer, while in entertainment environments conveying emotions becomes more important [29].

Several works in the literature have dealt with robot gesture generation from prosodic information; e.g. hand gesture generation using text, prosody, and dialogue act information in Android robots, [16], mapping of prosody cues to corresponding arm gestures for Nao robots [2,1], which later included head gestures [3,4], or gesture generation for storytelling humanoid robots considering both paragraph- and discourse-level prosody [14]. Other works have been devoted to achieve a fine synchronization between speech and gesture for humanoid robots, such as in [28], or in [25], where a Honda humanoid is used, and part-of-speech information is used, although only to provide timing information of spoken works in a text-to-speech system.

3 Experimental Framework

Since experimenting with humanoid robots is costly and potentially dangerous—accidental falls or unexpected movements may damage the robot itself and/or its surroundings—we conducted our experiments using a simulation software, able to model the physics that control the robot limitations of movement and

speed. Both the technical specifications and the environment for gesture generation are described in this section.

3.1 Technical specifications

For the implementation of our model, we used REEM-C [3], a full-size biped humanoid robotics research platform developed by PAL Robotics. The robot weighs 80 kg and is 165 cm tall. It has 44 degrees of freedom, offering applications for walking, navigation, grasping, face and speech recognition, and running over Ubuntu Linux 12.04 LTS and is ROS Hydro and OROCOS compatible. The ROS (Robot Operating System [4] is an open-source framework for robot software development, which provides standard operating system services such as hardware abstraction, low-level device control, implementation of commonly used functionality, message-passing between processes, and package management.

Pyttsx [5] and eSpeak [6] constitute the speech processing framework. Ptttsx is a cross-platform Python wrapper for text-to-speech synthesis that relies on the default speech synthesiser on each OS. For Ubuntu, it uses eSpeak, a compact open source software speech synthesiser written in C which supports SSML (Speech Synthesis Markup Language) and HTML. The simulation framework utilized Rviz [7] and Gazebo [8]. Rviz is a 3D visualisation tool for ROS, while Gazebo is a robotics simulator used to create embedded applications for a robot without depending physically on the actual machine, completely integrated with ROS.

The robot was controlled using *Play motion* [9], *MoveIt!* [10], and *Joint Trajectory Controller*. Play motion is a tool used to play pre-recorded motions on ros_control compliant robots via a simple actionlib interface; *MoveIt!* is a state-of-the art software for mobile manipulation, incorporating the latest advances in motion planning, manipulation, 3D perception, kinematics, control and navigation, and *Joint Trajectory Controller* is a controller for executing joint-space trajectories on a group of joints. Finally, *Sayit* is the responsible of the motions encoding and offers the speech and gesture synchronization functionality. Figure 1 shows an overview of the main components that constitute our experimental framework.

3.2 Gesture generation

As the trajectory controller used by REEM-C does not offer continuous trajectories, the use of Hidden Markov models to parameterize the model had to be discarded. Thus, we decided to build an expert system based on rules, generating our gesture database. We focused on the use of *beat* gestures, related to the prosody of speech, and *emblematic* gestures, which can just relate to the meaning of one single word, instead of relating to the meaning of a sentence or a group of sentences. The experimental gestures database, detailed in Table 1, was built by examining different TED talks [11] [12] and copying common gestures. For the beat gestures database, we selected 13 of the most common gestures the speakers performed, and we associated them with certain keywords (i.e. we associate the gestures *Nod* with the keywords *Hello*, *Hi*,

[3] http://pal-robotics.com/es/products/reem-c
[4] http://www.ros.org
[5] https://pyttsx.readthedocs.org/en/latest
[6] http://espeak.sourceforge.net
[7] http://sdk.rethinkrobotics.com/wiki/Rviz
[8] http://gazebosim.org
[9] http://wiki.ros.org/Robots/REEM-C/Tutorials/play_motion
[10] http://moveit.ros.org
[11] Julian Treasure: How to speak so that people want to listen, https://youtu.be/eIho2S0ZahI
[12] The power of seduction in our everyday lives, https://youtu.be/TBIL2sdfoVc

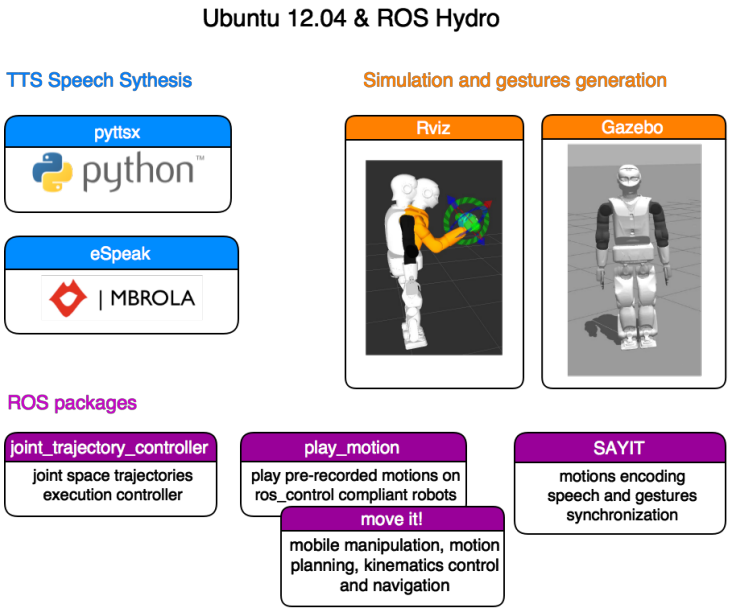

Fig. 1 Experimental framework overview. Top-left modules correspond to the speech synthesis software, responsible of the TTS and voice control. Top-right modules correspond to gestures design and simulation software. Bottom modules correspond to ROS packages involved in the process: movement control and speech and gesture synchronization.

Goodbye and *Bye*).For the emblematic gestures database, we selected 9 easily identifiable gestures [13]. Only gestures involving the upper part of the body (waist, arms, hands and head) were included due to REEM-C stability issues with legs spontaneous movement.

Gestures were manually designed with Rviz, a 3D visualization tool for ROS, and the *Joint states group grabber* Python module, which allows moving the robot joints in Rviz and grabbing the joints positions. Each gesture is encoded in a YAML file, a data serialization standard, and it is composed by a name, a list of the involved joints, and a list time points and joint positions that conform the gesture itself. Figure 2 illustrates the process of generating one of the gestures copied from a TED talk by imitating it in Rviz manually moving the robot's joints and getting their positions (blue circles).

4 Gesture and Speech Synchronization Approaches

This section presents the three different approaches to the problem that were studied along this work, namely: (a) part-of-speech (PoS)-based approach, (b) prosody-based approach, and (c) a combination of both PoS and prosody. Demonstration videos for each approach can be found online[14] [15] [16].

[13] A demonstration video with a display of all the beat and emblematic gestures can be found here: `https://youtu.be/G7XCf9LO66A`

[14] PoS-based approach demonstration: `https://youtu.be/KY-oLcJFV7U`

[15] Prosody-based approach demonstration: `https://youtu.be/zhjD5zWswws`

[16] Combined approach demonstration: `https://youtu.be/uNVkkxFsWfE`

Gesture	Description	Meaning
Nod	The head is tilted once up and down along the sagittal plane.	Agreement, acceptance, or acknowledgment.
Quick nod	The head is tilted once up and down along the sagittal plane, in a quicker and shorter move.	Agreement, acceptance, or acknowledgment.
Affirmation	The head is tilted in alternating up and down arcs along the sagittal plane.	Agreement, acceptance, or acknowledgment.
Head shake	The head is turned left and right along the transverse plane repeatedly in quick succession.	Disagreement, denial, or rejection.
Wave	The hand is raised and moved from side to side, palm facing outwards.	Greet someone, say goodbye.
Quick wave	The arm is raised and lowered again, palm facing outwards.	Greet someone, say goodbye or merely acknowledge another's presence.
Myself	A person points to himself, moving the hand towards the chest.	What is being said relates to himself or it is a personal opinion.
You	A person points to the person with whom he is talking.	What is being said relates to the other person, or an interaction is expected.
Here	The arm moves up and down while the index finger points out to the floor.	What is being said relates to the current location.

Table 1 Description and meaning of the nine gestures included in our emblematic gesture database. A demonstration video with a display of all the beat and emblematic gestures can be found here: `https://youtu.be/G7XCf9LO66A`

Fig. 2 Gestures imitation from TED talks on YouTube using Rviz. On the left side a frame of the original YouTube video. On the right side the imitation of the gesture with REEM-C in Rviz. The blue circles correspond to the robot's joints, manually moved to match the pose of the speaker.

4.1 Part-of-Speech (PoS)-based approach

Inspired by the *semantic and pragmatic synchronisation rules* suggested by [21], which state that co-occurring gestures and speech relate to the same idea unit and to the same pragmatic function, we used a shallow parsing technique to analyse the input text, identify certain keywords and fire events for the robot to perform gestures related to those keywords. We did not take into account the structure of the sentence

neither the role of the constituents in the main sentence.

The implementation of this event-driven module is as follows. First, the text to pronounce is stored in a SSML file (Speech Synthesis Markup Language), which allows specifying the speech rate, the voice to be used by the synthesizer, pauses, etc. Second, we scan the SSML looking for keywords related to our emblematic gestures. The SSML file is read, and a motion sequence is computed, assigning an emblematic gesture to each keyword found on the text, and a beat gesture to each content-word with no associated keyword. Content and non-content words are distinguished as a simplification of the morphological layer, as the controller can not manage too many gestures at the same time and they accumulate in a pipe, loosing all synchronisation with the text. Finally, for each uttered word, an event is fired by Pyttsx with the word that is about to be spoken. If that word is in our motion sequence, we command the robot to perform the appropriate gesture. When an emblematic gesture is fired, all gestures previously sent are canceled. Figure 3 shows the architecture used in this first approach.[17]

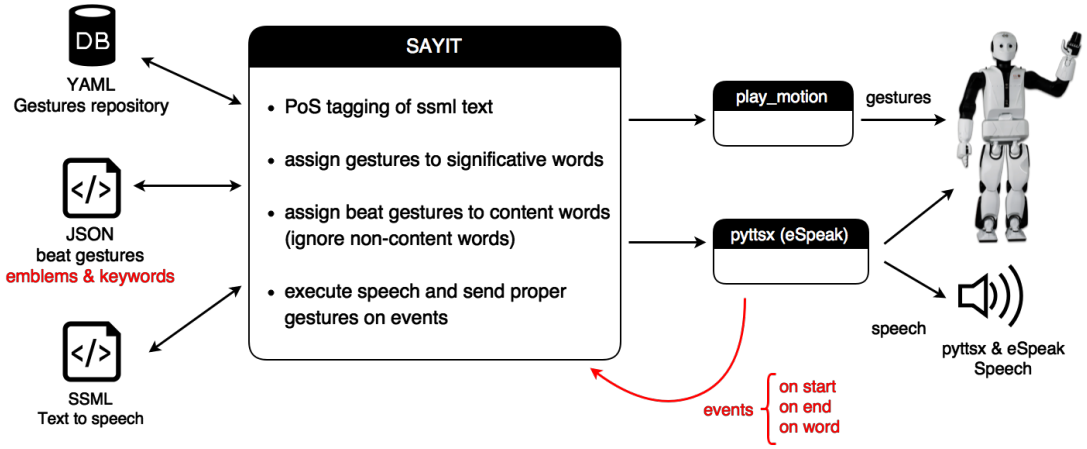

Fig. 3 PoS-based approach architecture. A motion sequence is computed assigning an emblematic gesture to each keyword found on the text and a beat gesture to each content-word with no associated keyword.

4.2 Prosody-based approach

Inspired by the *phonological synchronization rule* suggested by [21], which predicts that a gesture stroke should occur before the most prominent syllable, and according to all the literature being conducted that states that gestures and speech can be synchronized through prosody cues, we designed a prosody-based approach to synchronize our gestures with the speech. In this approach, we decided to use just beat gestures, instead of mixing beat and emblematic gestures, to better appreciate the difference.

Speech prosody is conveyed by the intonation, rhythm, and stress elements. In this work, we used the pitch value to model intonation. The goal was to align beat gestures with speech prosody, more precisely with the pitch. The main idea was to assign each beat gesture a certain *prosody curve*, in order to find the beat gestures sequence that better matches the speech prosody. For each beat gesture, we stored the name

[17] A demonstration video can be found here: `https://youtu.be/KY-oLcJFV7U`

and a list of poses pitch and time points, as a representation of the *prosody* of the gesture. For example, a certain beat gesture where we move both hands up and down twice to highlight certain parts of our speech (e.g. "We should do it quick and well", where we want to highlight the words "quick" and "well"), could be described with the sequence of time points and *pitch* [(250, 3), (500, 9), (750, 3), (1000, 9)], where the time points 500 ms, 1000 ms correspond to the two hands reaching the lower space point, accompanying the words "quick" and "well").

We used mbrola voices[18] to be able to extract the pitch and time points pairs from the SSML text. Figure 4 shows a plot of the pitch values for the demo text *Hello! My name is Reem and I am here to show you that it is possible to use a prosody-based approach to speech and gesture synchronization. As you see, I will keep moving while I talk according to the pitch accent of the text. Have a nice day! Bye!*. Then, we found the pitch peaks and assigned to each peak the gesture with the most similar pitch graph, randomly choosing if different motions were similar enough, and cleaned that sequence to delete gestures that overlapped with other gestures.

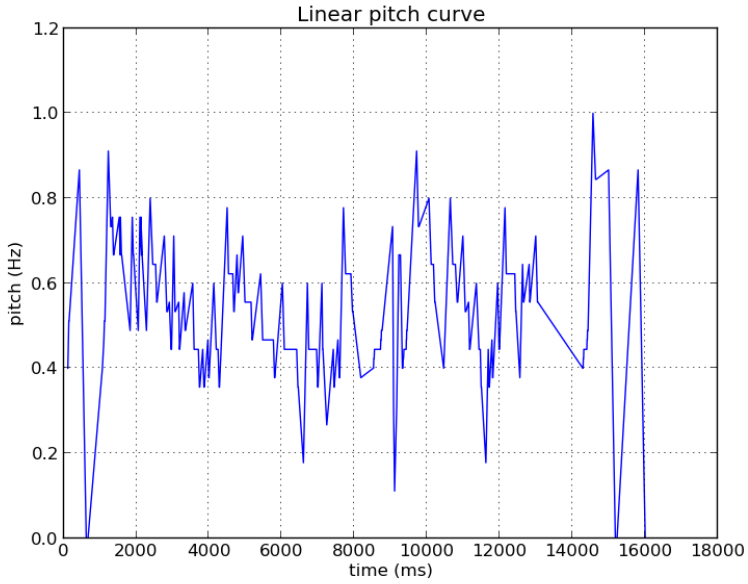

Fig. 4 Plot of the pitch values for the demo text *Hello! My name is Reem and I am here to show you that it is possible to use a prosody-based approach to speech and gesture synchronization. As you see, I will keep moving while I talk according to the pitch accent of the text. Have a nice day! Bye!*

Instead of working with events, as we had the time points, we launched the speech and sent to the robot the gestures in the right time point, using a timer. Figure 5 shows the architecture used in this approach.[19]

[18] `http://tcts.fpms.ac.be/synthesis/mbrola.html`

[19] A demonstration video can be found here: `https://youtu.be/zhjD5zWswws`

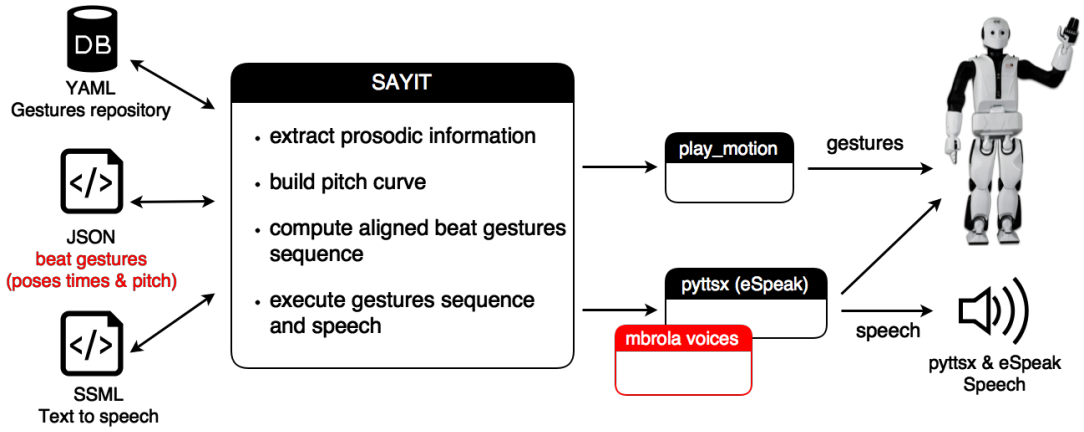

Fig. 5 Prosody-based approach architecture. A *prosody curve* is manually described for each beat gesture in the gestures database. Then, a motion sequence is computed by assigning the gesture with the most similar form to each pitch peak of the text to be spoken. No emblem gestures are performed.

4.3 Combined approach

As the prosody-driven approach seemed to offer a better gesture and speech synchronization—as we will discuss in the next section—we decided to develop a third approach combining the former two. The idea was to use the technique developed in the first approach to fire emblematic gestures, and use the prosody-based approach to perform appropriate and synchronized beat gestures.

This approach works exactly as explained for the prosody-based approach, with one difference: when the synthesizer is about to pronounce words associated with keywords, an event is fired and the proper emblematic gesture is sent to the robot to perform, as we did for the first approach. As emblematic gestures have specific meaning and are consciously used, we decided to prioritize them in front of beat gestures: while an emblematic gesture is performed, just another emblematic gesture can stop it and beat gestures supposed to be reproduced meanwhile are lost. Figure 6 shows the architecture used in this combined approach.[20]

5 Evaluation

The evaluation of our experiments is inherently difficult, as not only there are multiple correct gestures sequences for a given utterance, but appropriateness and naturalness estimation is highly subjective and only determinable by human observation. Thus, we conducted an evaluation survey to evaluate each of the three approaches presented in this work, where participants were presented three videos of the simulated REEM-C robot giving a short speech (around 30 seconds each) and gesturing according one of the three approaches, and were asked to evaluate them individually and by pairs. Table 2 summarizes relevant personal data of the participants. A total of 50 participants took part in the evaluation, 4% of them between 18 and 25 years old, 62% between 25 and 35 years old, 30% between 35 and 45 years old, and 4% over 45 years old. Regarding their English language level, 66% of the participants had a proficient level, 28% had intermediate

[20] A demonstration video can be found here: `https://youtu.be/uNVkkxFsWfE`

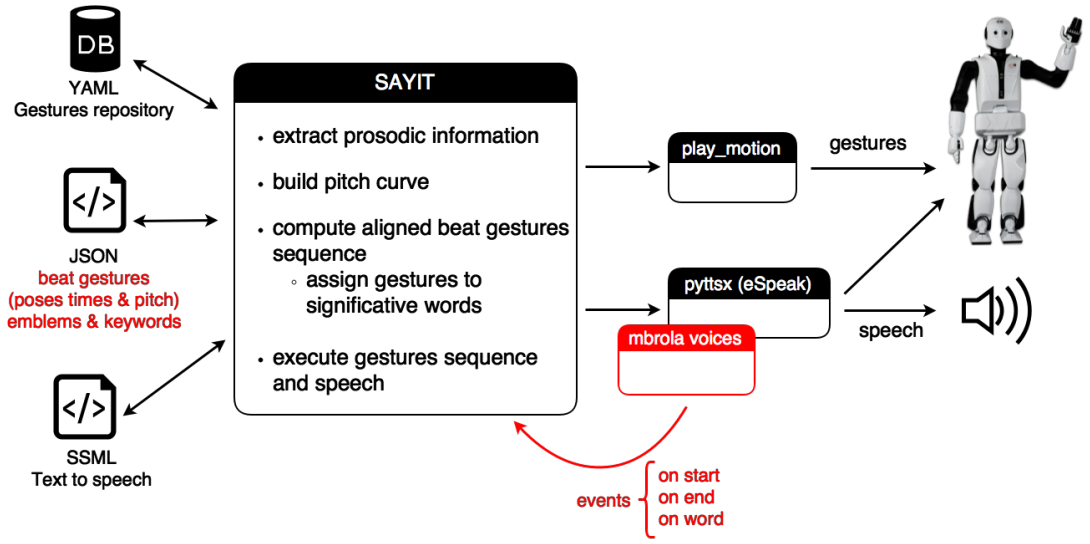

Fig. 6 Combined approach architecture. A *prosody curve* is manually described for each beat gesture in the gestures database. Then, a motion sequence is computed by assigning an emblematic gesture to each keyword found on the text, and assigning the gesture with the most similar form to each pitch peak of the text to be spoken.

knowledge and 8% declared to have basic knowledge. About their relation with speech technologies, 8% of the participants declared to be experts, 46% were familiar with them, and 46% had no related knowledge.

Age ranges	18-25 (4%), 25-35 (62%), 35-45 (30%), over 45 (4%)
English level	proficient (66%), intermediate (28%), basic (8%)
ST relation	expert (8%), familiar with them (46%), no related knowledge (46%)

Table 2 Description of the MOS test participants ages, English language level and relation with speech technologies. In parentheses the percentage over the 50 participants.

5.1 Pair-wise perception test

Participants were initially asked to compare the three approaches to each other by looking at pairs of videos corresponding to the PoS-based approach, the prosody-based approach and the combined PoS and prosody approach without any further information about which were the approaches being compared at each time. They were asked to compare them in terms of timing, appropriateness and general impression, choosing for each property which one they liked better according to the following definitions:

- Timing. The gestures are timed appropriately with the speech, even if the gesture meaning does not relate to the uttered words.
- Appropriateness. The performed gestures make sense, they relate well to the uttered words.
- General impression. The overall result looks natural.

Results show that PoS-based approach beats both the combined approach (66% vs. 34% in timing, 68% vs. 32% in appropriateness and 70% vs. 30% in naturalness) and the prosody approach (72% vs. 28% in

timing, 86% vs. 14% in appropriateness and 82% vs. 18% in naturalness), while the combined approach beats in turn the prosody-based approach (72% vs. 28% in timing, 98% vs. 2% in appropriateness and 90% vs. 10% in naturalness), as summarized in Table 3.

Test	Timing	Appropriateness	Naturalness
Prosody / Combined	28%/72%	2%/98%	10%/90%
PoS / Combined	66%/34%	68%/32%	70%/30%
Prosody / PoS	28%/72%	14%/86%	18%/82%

Table 3 Results of the pair-wise comparison. Percentages express participants preference for the first or the second approach being compared.

5.2 MOS perception test

After the pair-wise comparison, participants were asked to evaluate each approach individually by performing a mean opinion score (MOS) test, rating from 1 to 5 the timing, the appropriateness and the naturalness in the same terms as in the pairwise comparison, being 1 the worst score and 5 the best one. Again, results show that the PoS-based approach outperforms the other two in all three metrics, getting an average of 3.6 in timing, 3.8 in appropriateness and 3.7 in naturalness, while prosody gets 3.2, 2.4 and 2.5 respectively, and the combined approach 3.2, 3.4 and 3.3, as summarized in Table 4. Interestingly, we observe that the youngest participants, those between 18 and 25 years old, rated around 0.5 points higher both appropriateness and naturalness for all three approaches, and participants with no related knowledge in speech technologies rated the timing around 0.2 higher and the naturalness around 0.1 higher than participants with some previous knowledge.

Test	Timing	Appropriateness	Naturalness
PoS	3.6	3.8	3.7
Prosody	3.2	2.4	2.5
Combined	3.2	3.4	3.3

Table 4 Results of the individual evaluation of the three approaches. Mean scores on a 5 point scale.

6 Discussion and future work

We have presented three different approaches to extend humanoid robots communicative behaviour with upper-body gestures automatically synchronized with the speech by exploting prosody cues, part-of-speech and a combination of both, and we have proposed a method to characterize, design and store human gestures so they can be performed by REEM-C.

Regarding the gesture characterization, design and storing, we have described a classification of human gestures in five different types: emblematic, metaphoric, deictic, iconic and beat gestures. The first four relate closely to semantics, the study of the meaning, while beat gestures relate to morphology, which studies the structure of the language's morphemes and other linguistic units. We decided to use a simplified approach consisting of beat and emblematic gestures, discarding metaphoric, deictic and iconic gestures

that require a deeper analysis of the semantics of the input text. Manually designing the gestures using Rviz was very convenient, as it was simple and allowed us to generate as many gestures as needed, but it has two main handicaps: the resulting gestures do not look smooth, as we work with joint trajectories instead of with continuous trajectories, and the resulting YAML files are hard to read and difficult to fine tune.

The MOS test results show that the PoS and the combined approaches, which combined beat and emblematic gestures, clearly outperform the prosody-based approach that lacks emblematic gestures, indicating that those gestures related to the semantics of the speech contribute enormously to the appropriateness and naturalness perception. All three approaches present the same handicap: as Pyttsx fires the event when a word is about to be spoken it takes some time for the gestures controller to send a motion to the robot, and a small delay between the word and the gesture is clearly visible. Furthermore, for the two approaches using emblematic gestures, when an emblematic gesture is sent to the robot all current gestures are cancelled and the robot trembles due to the sudden change of gesture.

In the part-of-speech (PoS) approach, emblems are fired for each keyword, and beat gestures are fired for each content word. Beat gestures are queued and reproduced one after the other without cancelling, not adapted to the duration of the word they are supposed to accompany, and they eventually get desynchronized. However, this effect is not obvious to notice, and has little impact in final evaluation. In the prosody-based approach we use the pitch curve to model part of the prosody, enabling us to adjust our gestures *pitch* with the speech pitch. Even though it should result in a better timed gesture sequence, the lack of emblematic gestures and the empty time frames resulting of the cleaning of the overlapping gestures in the gesture sequence penalizes enormously the final result. Besides, describing the gestures *pitch* is difficult and very subjective. The third approach combines emblematic and beat gestures, firing emblems when keywords were to be spoken, and aligning beats with the prosody of the speech. The presence of both types of gestures contributes hugely to the final result, but the problems derived of the framework itself along with the problems described for the prosody-based approach are still present, deriving in a performance worse than expected.

Currently, PAL Robotics is developing a new whole body control module that will work with continue trajectories in the Cartesian space instead of working with joint trajectories. This will offer better stability and the possibility of launching multiple tasks at the same time, solving some of the problems encountered along this work, like the trembling effect when cancelling motions. Furthermore, it will allow researchers and developers to use hidden Markov models to parameterize the model. However, it will add other problems. For example, as it is being designed to maximize safety (of the robot and its surroundings), it will modify certain gestures in order to avoid the robot to fall. Incorporating the new controller would be the first thing to do in the future, taking advantage of all the functionalities it offers, but there are also some things to be improved in the module presented in this work. Incorporating a semantic module able to analyze, extract and exploit deeper meanings of the text, in order to select and use metaphoric, deictic and iconic gestures along with emblematic and beat gestures would certainly improve performance and would have a deeper impact in listeners. Moreover, designing more and more accurate gestures and improving the algorithm that computes the gestures sequence to minimize the empty time periods would also improve the obtained results.

7 Acknowledgements

The second author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE,

UE). The authors would like to thank the anonymous reviewers that helped to improve this paper through their valuable comments.

References

1. Aly, A., Tapus, A.: An integrated model of speech to arm gestures mapping in human-robot interaction. In: Information Control Problems in Manufacturing, vol. 14, pp. 817–822 (2012)
2. Aly, A., Tapus, A.: Prosody-driven robot arm gestures generation in human-robot interaction. In: Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pp. 257–258. ACM (2012)
3. Aly, A., Tapus, A.: Speech to head gesture mapping in multimodal human-robot interaction. In: Service Orientation in Holonic and Multi-Agent Manufacturing Control, pp. 183–196. Springer (2012)
4. Aly, A., Tapus, A.: Prosody-based adaptive metaphoric head and arm gestures synthesis in human robot interaction. In: 2013 16th International Conference on Advanced Robotics (ICAR), pp. 1–8. IEEE (2013)
5. Breckinridge Church, R., Garber, P., Rogalski, K.: The role of gesture in memory and social communication. Gesture **7**(2), 137–158 (2007)
6. Bremner, P., Leonards, U.: Speech and Gesture Emphasis Effects For Robotic and Human Communicators - a Direct Comparison. In: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, pp. 255–262. ACM (2015)
7. Bremner, P., Pipe, A.G., Fraser, M., Subramanian, S., Melhuish, C.: Beat gesture generation rules for human-robot interaction. In: RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 1029–1034. IEEE (2009)
8. Bremner, P., Pipe, A.G., Melhuish, C., Fraser, M., Subramanian, S.: The effects of robot-performed co-verbal gesture on listener behaviour. In: 2011 11th IEEE-RAS International Conference on Humanoid Robots, pp. 458–465. IEEE (2011)
9. Chiu, C.C., Marsella, S.: How to train your avatar: a data driven approach to gesture generation pp. 127–140 (2011)
10. Chiu, C.C., Marsella, S.: Gesture generation with low-dimensional embeddings pp. 781–788 (2014)
11. Ding, Y., Pelachaud, C., Artières, T.: Modeling multimodal behaviors from speech prosody. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8108 LNAI**, 217–228 (2013)
12. Fernández-Baena, A., Montaño, R., Antonijoan, M., Roversi, A., Miralles, D., Alías, F.: Gesture synthesis adapted to speech emphasis. Speech Communication **57**, 331–350 (2014)
13. Feyereisen, P., De Lannoy, J.D.: Gestures and speech: Psychological investigations. Cambridge University Press (1991)
14. Gelin, R., d'Alessandro, C., Le, Q.A., Deroo, O., Doukhan, D., Martin, J.C., Pelachaud, C., Rilliard, A., Rosset, S.: Towards a storytelling humanoid robot. In: 2010 AAAI Fall Symposium Series (2010)
15. Goldin-Meadow, S.: The role of gesture in communication and thinking. Trends in cognitive sciences **3**(11), 419–429 (1999)
16. Ishi, C.T., Machiyashiki, D., Mikata, R., Ishiguro, H.: A speech-driven hand gesture generation method and evaluation in android robots. IEEE Robotics and Automation Letters **3**(4), 3757–3764 (2018)
17. Kim, H.H., Ha, Y.S., Bien, Z., Park, K.H.: Gesture encoding and reproduction for human-robot interaction in text-to-gesture systems. Industrial Robot: An International Journal **39**(6), 551–563 (2012)
18. Krahmer, E., Swerts, M.: Audiovisual prosody — introduction to the special issue. Language and Speech (2009)
19. Le, Q.A., Hanoune, S., Pelachaud, C.: Design and implementation of an expressive gesture model for a humanoid robot. In: 2011 11th IEEE-RAS International Conference on Humanoid Robots, pp. 134–140. IEEE (2011)
20. Levine, S., Theobalt, C., Koltun, V.: Real-time prosody-driven synthesis of body language. In: ACM SIGGRAPH Asia 2009 papers on - SIGGRAPH Asia '09, vol. 28, p. 1. ACM Press, New York, New York, USA (2009)
21. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago Press (1992)
22. Meena, R., Jokinen, K., Wilcock, G.: Integration of gestures and speech in human-robot interaction. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pp. 673–678. IEEE (2012)
23. Mlakar, I., Kacic, Z., Rojc, M.: TTS-driven synthetic behaviour-generation model for artificial bodies. International Journal of Advanced Robotic Systems **10**, 1–20 (2013)
24. Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E.: Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological science **15**(2), 133–137 (2004)
25. Ng-Thow-Hing, V., Luo, P., Okita, S.: Synchronized gesture and speech production for humanoid robots. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4617–4624. IEEE (2010)
26. Nooteboom, S.: The prosody of speech: melody and rhythm. The handbook of phonetic sciences **5**, 640–673 (1997)
27. Pandzic, I.S., Forchheimer, R.: Mpeg-4 facial animation. The standard, implementation and applications. Chichester, England: John Wiley&Sons (2002)
28. Salem, M., Kopp, S., Joublin, F.: Generating finely synchronized gesture and speech for humanoid robots: a closed-loop approach. In: Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, pp. 219–220. IEEE Press (2013)

29. Tay, J., Veloso, M.: Modeling and composing gestures for human-robot interaction. In: 2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, pp. 107–112. IEEE (2012)
30. Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: An overview. Speech Communication **57**, 209–232 (2014)
31. Wennerstrom, A.: The music of everyday speech: Prosody and discourse analysis. Oxford University Press (2001)
32. Zoric, G., Forchheimer, R., Pandzic, I.S.: On creating multimodal virtual humans—real time speech driven facial gesturing. Multimedia Tools and Applications **54**(1), 165–179 (2010)