Fernández De Gorostiza Luengo, Javier; Alonso Martín, Fernando; Castro-González, Álvaro; Salichs, Miguel Ángel (2017) Sound Synthesis for Communicating Nonverbal Expressive Cues. *IEEE Access,* v.5, pp.: 1941-1957.

DOI: https://doi.org/10.1109/ACCESS.2017.2658726

Sound Synthesis for Communicating Nonverbal Expressive Cues

Javi F. Gorostiza, F. Alonso-Martín, A. Castro-González, and M. Salichs

***Abstract*—Non-verbal sounds (NVS) constitute an appealing communicative channel for transmitting a message during a dialog. They provide two main benefits: they are not linked to any particular language, and they can express a message in a short time. NVS have been successfully used in robotics, cell phones, and science fiction films. However, there is a lack of deep studies on how to model NVS. For instance, most of the systems for NVS expression are *ad hoc* solutions that focus on the communication of the most prominent emotion. Only a small number of papers have proposed a more general model or dealt directly with the expression of pure communicative acts, such as affirmation, denial, or greeting. In this paper we propose a system, referred to as the Sonic Expression System (SES), that is able to generate NVS on the fly by adapting the sound to the context of the interaction. The system is designed to be used by social robots while conducting human–robot interactions. It is based on a model that includes several acoustic features from the amplitude, frequency, and time spaces. In order to evaluate the capabilities of the system, nine categories of communicative acts were created. By means of an online questionnaire, 51 participants classified the utterances according to their meaning, such as agreement, hesitation, denial, hush, question, summon, encouragement, greetings, and laughing. The results showed how very different NVS generated by our SES can be used for communicating.**

***Index Terms*—Sound synthesis, Human–Robot Interaction, Electrosonic mode, Social Robots, Non-Verbal Sounds, Sonic mode, Quasons**

I. Introduction

One of the most important challenges that social robots have to face is to be able of interacting in a relatively natural, efficient, and coherent way. A natural interaction implies many robot capabilities that include multimodality, adaptability, co-operativeness, reactivity, and others [**?**]. There are some papers that deal with gesture expression or gesture perception [**?**], [**?**]. Also, in voice interaction (dialog management), there are many papers tackling problems such as grounding [**?**], engagement [**?**], natural language understanding, and natural language generation [**?**]. Much ongoing research also treats multimodality on both sides: perception and expression. For instance, [**?**] presents a multimodal system that is able to combine the inputs from a tablet, the user's voice, and gestures. Multimodal fusion allows completing the information of one mode by using another, for example, the system resolves deixis cases in commands such as "go there" (while pointing to some place with the hand). In [**?**], a multimodal dialog manager allows a robot to combine partial information of vision and speech into a coherent message. Many models, such as Behavior Markup Language (BML), face the problem of how to synchronize the different modalities of a robot or agent into a unique coherent expression [**?**].

However, there has been little research that analyzes the communicative possibilities offered by sound expression to robots, i.e., the "sonic mode."

The sonic mode is a constant channel present in human activity, which includes the voice, but is not limited to it. Humans, and of course other animals, communicate with each other by: non-linguistic utterances, suprasegmental sounds, sounds such as laughing, sighing, or yawning, and many others that involve different parts of the body (clapping or intestinal sounds, for example). Sounds and varieties of vocal patterns (e.g., a song) play an important role in social relationships. As noted by Bruce Richman [**?**], a prosodic variation while vocalizing relaxes inner tensions and conflicts in the background of social groups, and "succeeds like social grooming in setting up minute-to-minute relationships." This "sonic mode" includes different types of NVS: "non-linguistic utterance" (NLU) in [**?**], "subtle expressions" in [**?**], "gibberish speech" in [**?**], "musical sound" in [**?**], "anthropomorphic auditory icons" in [**?**], "affect burst" in [**?**], or "auditory icons" in [**?**].

Also, from a cognitive perspective, non-linguistic elements are quite important in artificial intelligence agents for connecting language to the real world [**?**]. Language rests upon deep roots that are not verbal but acoustic. These non-verbal contexts leverage speech acts, words, and utterances, to convey their meaning.

New electronic devices, such as mobile phones, tablets, and computers, use the sonic mode to convey relevant and specific events, such as the reception of a new message, a new incoming call, low battery, etc. These sounds, called *earcons*, are composed in such a manner that they analogously imitate the event they are intended to express, and are easy to understand. For instance, a low battery earcon could decrease its *pitch* to express that the battery charge is in fact decreasing.

Similar to these devices, social robots are essentially electronic devices that facilitate the task of generating, creating, and expressing electronic complex sounds. A robot can reproduce music or any natural sampled sound, but it can also express itself by artificial electronic sounds similar to the ones generated by other electronic devices, such as music synthesizers. Sonic design is an appealing research field—as pointed out in [**?**], "Voice and gesture provide a rich domain for sketching which is just waiting for the appropriate tools that can be exploited for sonic interaction design."

The **Sonic Expression System** (SES) is a synthesizer system that allows a robot to express itself by generating different electronic nonverbal sounds (NVSs) in real time. This system may be desirable for natural human–robot interaction

Manuscript received; revised .

(HRI). The sonic mode comes in as a new appealing robotic expression mode. This fact implies several new challenges. For instance, how to control electronic sounds for expressing emotions, intentions, or communicative acts such as the signals of greeting, affirmation, or non-understanding.

Several papers on NVS generation focus on the expression of the most prominent emotions, such as anger/irritation, fear, disgust/dislike, happiness/joy, sadness, surprise, sorrow, neutral/calm, comfort, distress, shyness, pride, or expectation (see [**?**], [**?**], [**?**], [**?**]). But only a few deal directly with the expression of pure communicative expressions or intentions, such as affirmation/agreement, denial/disagreement, encouragement, introduction, questioning, or hesitation, as described by Silbot in [**?**], or the "subtle expressions" described in [**?**].

This shortcoming reflects the difficulty in expressing communicative acts by NVS with the systems developed at the moment.

This paper not only proposes a general model for NVS generation, but also develops a particular SES, based on such model, that focuses on those types of communicative intentions rather than on the expression of emotions. The developed system is tested in expressing the following intentions or communicative acts: agreement, hesitation, denial, question, hush, summon, encouragement, greeting, and laughter.

The sonic mode is seen as a complementary modality in a multimodal interactive system, and not just as an alternative solution in either natural HRI or in human–computer interaction (HCI). We believe that it will enhance the expressiveness, eloquence and efficiency of the interaction with a human being.

Even though natural spoken language, composed of utterances, includes hesitations, pauses, non-verbal sounds, repetitions, and in essence contains non-structural elements, every utterance has an ordered structure formed by simpler sound elements, such as phonemes and other guttural sounds combined with moments of silence. Therefore, SES allows controlling these basic elements or *grains*, and must be versatile enough to be adaptable online at interaction time.

A complete SES should be able to express a variety of possible sound domains: nature and objects such as waterfall, a thunder or a glass crashing; human sounds such as clapping, yawing or laughing; artificial sounds such as the ones used by *earcons*, sci-fi or robotic sounds. Each of these sounds has different technical requirements.

The paradigm for sonic synthesis followed by the presented system is a special type of granular synthesis where each grain incorporates multiple acoustic features that can be modified on-line. The rules of combination of the grains into more complex synthesized sounds are also addressed in the present paper. **Each grain is called a *quason*, and their meaningful combination forms a *Sonic Utterance* (SU).** Both concepts will be described in more detail below.

A. Requirements for a General Sonic Expression System

The main purpose of our project is to create a general and adaptable Sonic Expression System (SES), as defined above, that could express sonic utterances in four main groups: intentions or communicative expressions, such as approval, rejection, hesitation, or greeting; expressing affection, such as joy, calmness, or sadness; human nature sounds that are different from an explicit communicative expression, such as laughter, weeping, coughing, yawning, or a heartbeat; and finally, narrative communication messages by pantomimic sounds that represent a particular occurrence or event, such as something that falls or breaks, something that suddenly happens, or something that goes away. We consider that these four groups cover most of the communicative necessities for many HRI scenes.

Natural interaction implies an adaptation and synchronization between the involved subjects. For example, humans tend to adapt their rhythms both in gesture and voices. Message synchronization plays a significant role in the naturalness of an interaction, as has been pointed out by Kendom [**?**], Birdwhistell [**?**], and many others (a good compilation of papers that deal with the rhythms of interaction can be found in [**?**]).

The expression of non-verbal sounds is interactive and has to adapt the emitted sound to fit a particular communicative circumstance in the interaction loop. For instance, instead of having pre-recorded sounds, real-time sound synthesis allows defining the appropriate duration of an NVS in order to fit correctly into the latent rhythm of the messages in a natural interaction.

This synchronization implies both adapting the timing and modulating the intensity. For instance, in order to express a sound of surprise, different acoustic features have to be modulated depending on the importance, intensity, and energy of the surprise, which will go from a subtle "uhm?" to an energetic "whaaaat!!!" Other sounds, such as expressing a greeting, may need their duration and timing adapted to fit the rhythm of the interaction.

The main requirements for a general SES could be summarized in the following:

- Expressivity. The sounds should cover a wide range of domains, from communicative intentions to pantomimic sounds.
- Adaptation and synchronization to the user's movements and sounds.
- Real-time synthesis. It should conform to the interaction context.

This paper is structured as follows.

First, in Section II, the state-of-the-art of non-verbal sound generation systems is presented. Next, in Section III, the quason concept is defined. The next section, Section IV, explains the combination of quasons in sonic utterance. Later, in Section V, the implementation of SES is described. In Section VI, several examples are shown to demonstrate the versatility of our system. Next, Section VII presents a study that allows checking whether the expression of the non-verbal sounds by our system is potentially communicative, so it can enhance and improve natural HRI. Lastly, Section VIII presents our conclusions and outlines future research.

II. Related Work

At a glance, in the state-of-the-art there is no general formalization of an acoustic control model for generating NVS. Each researcher uses a different model and acoustic parameters. Moreover, each set of parameters is closely related to the model and the algorithm implemented for generating the NVS. The main purpose of this section is to establish the basis of a general model that includes all the parameters of the most relevant SES paradigms.

Several researchers, such as Juslin et al. [**?**] and Coutinho et al.[**?**], have analyzed the relationship between the psychoacoustic features of music and the emotional feelings evoked in the user. For instance, [**?**] presents a hybrid model that gathers the acoustic features in both speech and in music.

Table I summarizes some prominent SES models/parameters: a human voice model developed by [**?**] and used by Kismet, the robotic torso ([**?**]); the system presented in [**?**] that simulates how to synthesize gibberish for emotional agents; and the main parameters used for studying the detection of emotions in music in [**?**].

Each model is represented by a finite set of acoustic parameters. These parameters are gathered into three main acoustic categories: Amplitude, Frequency and Time.

Amplitude includes the parameters that directly affect the changes over time in the signal's energy. Frequency includes all the parameters that affect the spectrum of the signal. The parameters included in the time category affect the duration of the whole signal or a part of it. Each model depicted in Table I is explained in more detail in the following sections.

This section also mentions the NVS generation systems used in robotics.

A. Prosodic Features of Human Voice Models

The acoustic utterance models for expressing emotions through speech are based on phonetic and syntactic parameters. The first column of Table I is related to the general acoustical model described in [**?**], which was applied to the Kismet robot, as described in [**?**] and [**?**].

There are some parameters with a direct influence on the control of the *amplitude envelope* of the final utterance. For instance, accents and emotive emphasis are made by increasing the volume in a certain part of the utterance, which is controlled by the *accent_shape* and *contour_slope* parameters. The general volume is controlled by *loudness*. The parameter *tremor* controls the irregularities between successive glottal pulses.

Parameters such as *breathing*, *brilliance* or *laryngealization* have a specific influence on the *timbre* of the utterance. Thus, breath includes pink noise in the signal, and brilliance is achieved by a HPF (high pass filter). Pitch variation is described by parameters such as *average_pitch*, *pitch_range*, *pitch_base* and *pitch_discontinuity*.

Timing aspects are covered by the *precision_of_articulation* parameter and the *speech_rate* parameter (the velocity of the sound signal).

These parameters are continuous and normalized in the range of $[-10, 10]$ (minimum and maximum influence); zero is regulated to be the neutral influence. The system takes as input an emotion label, from among 13 possibilities, and an incoming text sentence.

For each emotion label, there is a vector of fixed default values for the parameters that are adjusted by the developer. Going beyond or below these values, the influence or intensity will increase or decrease, respectively. This calibration is handmade by the developer.

B. The Generation of Emotional Speech in a Cartoon

In [**?**], the author describes an algorithm that allows an artificial agent to modulate its intonation to express emotions, concatenating speech synthesis as gibberish. Therefore, as in [**?**], the system also makes a correlation between an incoming emotion label and a set of values.

A simple and complete algorithm is in charge of specifying the pitch contour, and the duration, for each phoneme of the final utterance.

It is based on both continuous and logical parameters. For instance, *PROBACCENT* is a continuous parameter that defines a probability rate for stressing the phoneme. Other parameters, such as *CONTOURLASTWORD*, or *LASTWORDACCENTED*, just take two logical values that define whether the last phoneme pitch has to be brought up or let down by the amount set by the *PITCHVAR* parameter. Accents in this model are made by stretching the duration of the accented phoneme.

In order to give naturalness to the final sentence, some parameters such as *DURVAR*, the variation of the duration of the phoneme, and *PITCHVAR*, the variation of the pitch of the phoneme, are used inside a random function. Therefore, the algorithm, when executed different times but with the same values of the parameters, can generate acoustically different sentences.

There is one main parameter, VOLUME, that sets the loudness of the complete sentence.

C. Music and Emotions

In [**?**], the close relationship between the vocal and musical expression of emotions is demonstrated by reviewing 104 studies. Both channels (voice and music) reveal similarities in "the accuracy with which discrete emotions were communicated" and "the emotion-specific patterns of acoustic cues used to communicate each emotion" [**?**]. In that paper, some acoustic features were defined as cues for depicting where the essence of the emotion in the voice/music was: the fundamental frequency (F0), pauses, volume contour, rhythm, articulation, and speech rate. Some of them are more connotative and difficult to define in mathematical terms.

The expression of emotion in music as described in [**?**] is focused on the understanding of the specific psychoacoustic features involved in the expression of emotions in music and speech. That paper relates emotional parameters, arousal and valence to a large set of psychoacoustic parameters that are classified into five main categories: Dynamics, Loudness, Timbre, Mean Pitch, and Pitch Variation.

Human Voice Model (Kismet and DECtalk [?])	Gibberish (Oudeyer [?])	Music and Emotions (Coutinho [?])	Parameter Category
accent_shape *loudness* *contour_slope* *tremor*	*VOLUME*	*Dynamic Loudness*	**Amplitude**
average_pitch *final_lowering* *pitch_range* *reference_line* *breathiness* *brilliance* *laryngealization* *pause_discontinuity*	*MEANPITCH* *PITCHVAR* *MAXPITCH* *CONTOURLASTWORD* *DEFAULTCONTOUR*	*Melody Contour* *Prosody Contour* *Spectral Flux* *Sharpness (Aures)* *Sharpness (Zwicker & Fastl)* *Spectrum Centroid* *Dissonance* *Roughness*	**Frequency**
exaggeration *fluent pauses* *hesitation pauses* *speech rate* *stress frequency* *precision*	*LASTWORDACCENTED* *MEANDUR* *PROBACCENT* *DURVAR*	*Tempo* *Speech Rate*	**Time**

TABLE I: Acoustic parameters taken from the most relevant papers of the state-of-the-art

These psychoacoustic parameters are summarized in column three of Table I. Continuous parameters, such as *Melody/Prosody contour*, control the evolution of pitch along time for the contour of the prosody of the music/speech. The *spectral flux* in the case of music, and the *roughness* in the case of speech, quantify how much the power spectrum of the signal changes in time. The *Power Spectrum Centroid* and the sharpness of the acoustic signal in two mathematical definitions, one defined by Zwicher and Fastl (see [**?**]), and the other defined by Aures (see [**?**]), define the timbre of the signal, that is, the evolution of its spectrum over time.

Tempo, for music, and speech rate, for speech, set the rhythm velocity of the acoustical signal.

D. Sonic Expression Systems in Social Robots

NVS have been used extensively by several science fiction robotic characters, such as R2D2 and WALL-E. The sounds of these robots were designed by Ben Burtt, a sound designer who also made the sounds of Darth Vader breathing, the classical lightsaber hum, and many other important sci-fi sounds. He was able to efficiently combine natural sounds with electronic-sounding effects to produce NVS that expressed R2D2's intentions, thoughts, and emotions.

R2D2's sounds are rapid whistles whose pitches, which are quite high, jump quickly in patterns of rhythmical licks. Some of R2D2's sounds have been represented in [**?**] as musical scores. They have been analyzed in terms of musical parameters, such as intonation, pitch, and timbre. The analysis gives the general rules of the behavior of these parameters when the robot expresses five communicative intentions: affirmation, denial, encouragement, introduction, and question, and two main emotions: happiness and sadness. This result relates each of these seven expressions with a description of the used intonation, pitch range, and timbre. Thus, Affirmation is described as a "descending progression of short sixteenth note" with a pitch in the range of 262–1.175 Hz and using a timbre identified as a whistle and a synthesizer.

This analysis is quite connotative and qualitative, but can be used as a basic description of a more general model for those seven expressions.

These rules served as the inspiration for the composition of NVS as short fixed musical licks for the robot Silbot ([**?**]), in an application where the robot works as an English teacher [**?**]. Although this robot is able to express a set of five intentions and three emotions, it is not designed to be extrapolated to any other robot.

[**?**] studied how schoolchildren perceive different variations of musical parameters of the emotive aspect of NVS in a humanoid robot, *Nao*. The system generates simple tones without any possibility of controlling their timbre. It neither allows generating more than one tone at a time nor adding harmony to the final sound. As for the timing aspect, there is no rhythm, and the only parameter used to generate the sounds is their duration.

The sounds presented to each child had basic variations in pitch contours and duration, and the children had to identify the dominant emotion of what they heard.

The results suggest that variations in timing have more influence than variations in pitch contour.

In essence, the algorithms of the reviewed models in the previous section share the same mechanism: they modulate the general physical features of the acoustic utterance: the amplitude envelope, the frequency or pitch variation, the evolution of the signal spectrum over time, and the duration of the utterance. Despite the fact that different algorithms and sets of acoustic features produce similar expressive sounds, there is no unique general model or formalism for a general SES. Each paper uses its *ad hoc* SES solution.

In general, all the acoustic features presented can be gathered into three main categories: Amplitude, Frequency and Time, as depicted in Table I. Moreover, for simulating a live

expression, all the models use some type of randomness or probability parameter in the assignment of the values to some acoustic features.

III. The Description of the *Quason*: The Atomic Unit of Sound

The proposed model represents the sound landscape of a whole sonic expression as a musical combination of one or more minimum elements. A *quason* is a model that represents each of these indivisible, minimum sounds in terms of the variation of its acoustic features.

We define a quason as the smallest sound unit that holds a set of indivisible psychoacoustic features that makes it perfectly distinguishable from other sounds, and whose combinations generate a more complex individual sound unit. The name quason comes as a combination of the words *quantum*, in the sense of indivisible package of information, and *sound*.

A. The Acoustic Features of the Quason

The main acoustic features used for the synthesis of quasons are classified into three categories: Amplitude, Frequency and Time. The equivalent of a quason in classical music notation would be a simple note. Each note has its own graphical representation that describes its pitch and duration. Since quasons include additional parameters besides their fundamental frequency and duration, it is necessary to create a new graphical representation. Figure 1 shows these three categories, Ω_{amp}, Ω_{freq} and Ω_{time} in a new graphical representation for quasons.

1) The Parameters of the Quason in the Amplitude Space Ω_a: This category includes two parameters: the **amplitude envelope** and the **volume** of the quason. The amplitude envelope is defined to be a normalized curve (in the range between 0 and 1) as shown in Figure 1. The normalized amplitude is later scaled using the volume parameter. In Figure 1, an example of an amplitude envelope is represented by a set of 12 key-points.

2) Quason Parameters in the Frequency Space Ω_f: The category Frequency refers to the spectrum of the quason, which is defined by two curves: the **pitch envelope** and the **timbre envelope**.

The pitch envelope is the evolution in time of the quason's fundamental frequency while it is sounding. Analogously to the amplitude envelope, the pitch envelope is defined to be a normalized curve, which is scaled by two main parameters: the main pitch ($F0$), and a percentage frequency variation (ΔF).

In Figure 1 the pitch envelope is implemented as a nominal curve formed by six key-points represented by small squares, with its normalized value.

Another important acoustic feature in the frequency domain is the timbre. In psychoacoustics, timbre defines the tone color, personality, and tone quality ([**?**]). Physically, it is related to the spectrum of the sound.

The study of timbre is complicated enough by itself, but in our model, we simplify the timbre to be a uni-dimensional parameter. Its variation is defined by a normalized curve, the *timbre envelope*, where $timbre = 0$ represents a soft or dull timbre, with few harmonic components, and $timbre = 1$ represents a sharp or noisy timbre with more harmonic and non-harmonic components. How timbre is implemented is explained in more detail in Section V-B. In Figure 1, the timbre envelope is implemented as the variation of the width of the pitch envelope curve.

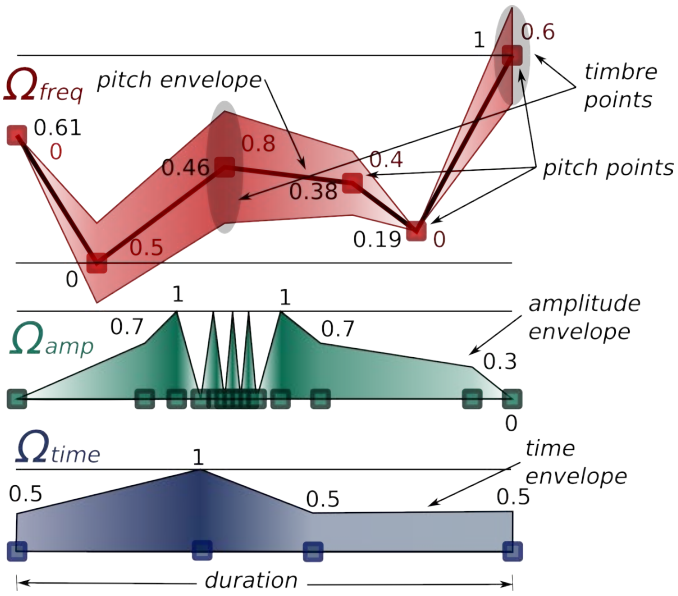

Fig. 1: Graphical representation of the quason concept, including its acoustic features classified into three categories: Amplitude, Frequency and Time. Each quason is represented by four main envelopes: amplitude, pitch, timbre and time. Each timbre envelope point has a scalar timbre value, which has been represented as the width of the pitch envelope curve. Notice that curves are described as a set of key points and they are normalized in the range [0,1]; the system interpolates between such points.

3) Quason Parameters in the Time space Ω_t: The main feature of a *quason* in this category is the **duration**, i.e., the time the quason is audible. This parameter defines the duration of the envelopes related to the amplitude, pitch and timbre, described above.

There is also a *tempo curve* that defines the variation as a percentage of the tempo, or the time velocity at each instant. For instance, in Figure 1, the tempo envelope of the quason is defined by a set of four key-points. It will begin at 50% of its normal value. Then the envelope will begin an *accelerando* to reach 100% of its normal velocity, and later will decrease again to 50% until the end of the quason.

Notice that the tempo curve does not affect the quason's pitch or other categorical feature.

IV. The Sonic Utterance: A Combination of Quasons

A *Sonic Utterance* (SU) is defined here as a sound structure formed by a combination of a finite number of quasons in a musical lick, that is, in a musical phrase with a communicative purpose.

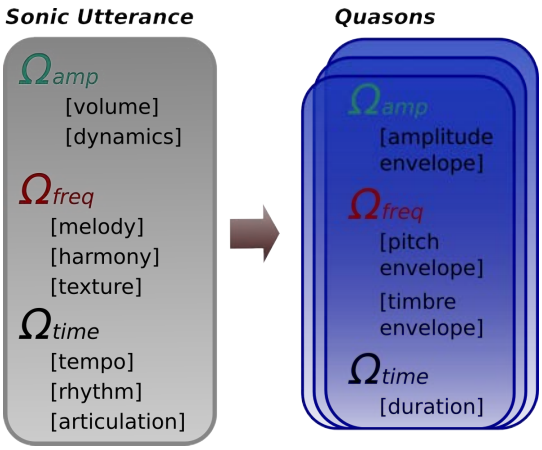

Fig. 2: Acoustic features of a Sonic Utterance as a set of quasons. The high level SU acoustic features modulate the low level acoustic features of the quasons.

A. Acoustic Features of Sonic Utterances

SU acoustic features are also defined in three categories: Amplitude, Frequency and Time, as shown in Figure 2. When the SU is played, its high level acoustic features modulate the low level features of the quasons, as explained below.

a) Amplitude Category (Ω_a): In the domain of amplitude, the SU defines the **global amplitude envelope**, independently of the particular local amplitude envelope of each quason.

The **Volume** parameter is also defined here as the global volume of the sonic utterance, and it is used to scale the global amplitude envelope curve, which is a normalized curve.

The combination of SU volume and SU amplitude envelopes scale the volume of each quason that belongs to the SU.

b) Frequency Category (Ω_f): **Melody** defines an ordered list of the main frequencies of each quason in the SU. The melody could belong to a specific map or musical scale if there is interest in giving a musical intention to the sonic utterance. For instance, as described in several papers, such as [**?**], minor scales are associated with the perception of negative messages, affects, or emotions, such as sadness, while a major scale tends to carry a more positive message, such as encouragement.

Several quasons could sound at the same time. **Texture** measures the multiplicity or number of quasons simultaneously present in the sound of the SU, which establishes a relation between their pitches that is called **harmony**. This relation could express relaxation, unison, and agreement, or it could express dissonance, discord and tension.

Having more than one voice line in the SU increases enormously the expressivity of the communicative sound. But it also brings up two main problems: how to harmonize the different voices, and how to combine them in time and in the rhythm of the utterance elements for communicating a non-verbal message. This aspect is analyzed in more detail by a few SU examples in Section VI.

c) Time Category (Ω_t): The sonic utterance is composed of different quasons following a **rhythm pattern**, which establishes the place in time where each quason has to sound.

This pattern is created at two levels: horizontally, following the definition of a melody, which concerns the quasons of the main musical voice line, and vertically, following the definition of harmony, which concerns the accompaniment quasons. For instance, a counterpoint based approach will have two independent voices in their complementary rhythm patterns, which will produce a different communicative intention than an SU where all quasons follow the same rhythm pattern.

Tempo establishes the velocity of the SU. It is measured in beats-per-minute (bpm). There is also a **tempo modulation** described by a nominal curve that specifies the main tempo modification along the sonic utterance.

The **articulation** parameter is the tempo variation for each quason. The rhythm can be mechanical and regular, or include variations, imperfections, and be more natural. A more natural SU includes an irregularity articulation parameter that sets a percentage of randomness in each pulse of the tempo.

The tempo establishes the duration of the sonic utterance that scales in time the duration of each quason.

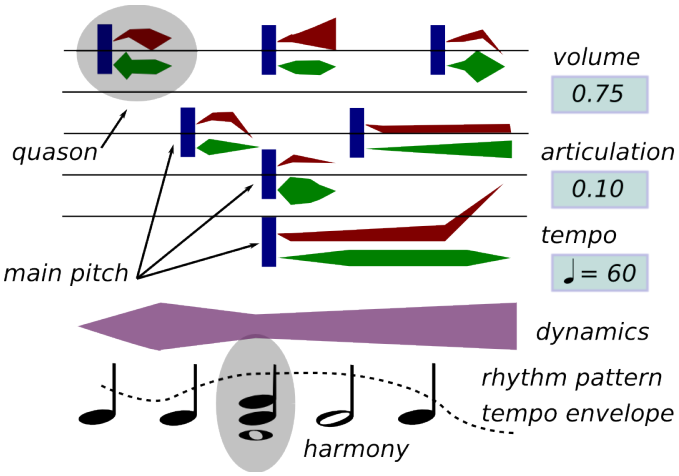

Fig. 3: Example of a sonic utterance with all its acoustic features. The quasons are combined in a musical score. The main amplitude envelope represents the musical dynamics as a normalized curved that is scaled by the main volume (dynamics). The position of each quason in the score establishes the rhythmic pattern of the SU, which is played as defined by the tempo parameter in beats per minute (bpm). The articulation establishes a random percentage of variation in the regularity of that tempo. The vertical position of the quason defines its main pitch, which will be used to scale the frequency envelope of each quason.

B. An Example of a Sonic Utterance

Figure 3 shows an example of an SU in a musical score composed of seven quasons. As in a classical score, the horizontal dimension represents time while the vertical dimension represents pitch. A **rhythm pattern** in the low part of the figure shows the rhythm figure for each quason. The **tempo** parameter establishes the velocity and the duration of the sonic phrase. Notice that some quasons can sound concurrently,

which causes a **harmonic** relationship, but there is also a **melodic** line that is generally established by the highest voice.

The height of the quason in the score is related to its main pitch, **F0**, which is used to scale its **pitch envelope**. The dynamic curve and the main **volume** modulate the amplitude of each quason of the SU.

The tempo is modulated by a **tempo envelope**, a nominal curve that establishes the percentage of modulation. There is also a feature, the **articulation** parameter, that includes a random percentage of variation in the tempo of playing the rhythmic pattern.

V. The Sonic Expression System

Figure 4 shows the complete SES with its two main modules. The SU, which is formed by quasons, is synthesized by the SU Player and the SU Controller.

The SU Controller sends, in real time, the acoustic parameters variations to the SU Player. The SU Player sends Open Control Sound (OSC)[1] messages to the implemented Puredata application[2], which synthesizes the sound in SES.

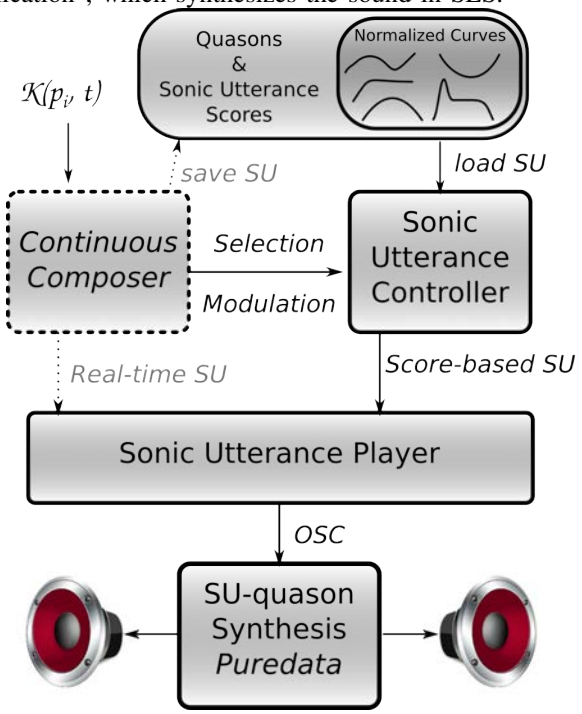

Fig. 4: Sonic Expression System (SES). SU-quasons can be generated in real-time by the Continuous Composer or from a modulation of SU-quasons of a repository. This paper focuses on the functionality of modules in continuous line.

There is an initial *quason repository* that includes the basic elements that can be modulated and combined into new quasons. Basic quasons are built from a set of normalized curves that represent the shape of an envelope, for amplitude, pitch, timbre or rhythm.

Both the initial sets of quasons and the normalized curves can be easily increased with new elements at interaction time. This is achieved by the Continuous Composer module, which creates normalized curves from different sources, as, for instance, from the acoustic features perception of the user's utterance. The details of such composition are outside the scope of the present paper, and what is here relevant is that this module receives the communicative act as a function of time of some articulation parameters $\kappa(p_i,t)$, such as the timing of the SU, its duration, or intensity, which can be used in the generation and modulation of the SU. In this manner, the SES allows modulating and adapting the SU to the context of the interaction: for instance, adapting the timings, rhythms, and pitches to adjust the robot's expression to the users.

Once the Continuous Composer module receives the communicative act, it selects the SU-quasons from the repository that is loaded and modulated by the SU Controller. Modulation changes the values of the acoustic parameters of both the SU and the quasons that make up that SU, as explained in Section V-A.

A. Nominal Curves and Modulation

Continuous parameters, such as amplitude, pitch, and timbre, are based on a repository of a set of nominal curves. These curves represent the shape of an envelope. For instance, a sound whose amplitude goes up will load the *up.env* nominal curve and then modulate it using the volume and duration parameter to scale the nominal curve for the amplitude envelope.

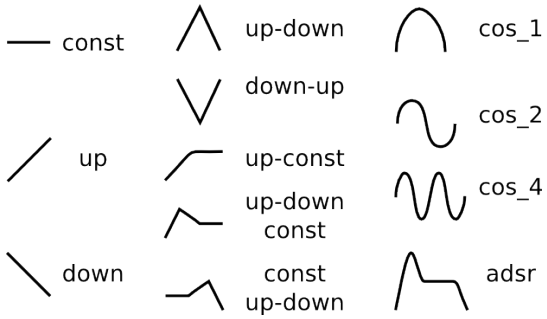

Fig. 5: Initial repository of normalized curves. These curves are used for making up quasons by selection and modulation. The repository can easily be increased at interaction time.

As shown in Figure 5, there are nominal curves for going up, down, or for the standard Attack-Decay-Sustain Release envelope (ADSR). Each curve can be assigned to any of the continuous parameters of the quason: amplitude, frequency, or timbre envelopes. In the definition of the quason, the amplitude range of each curve and the duration of the quason is also defined. These values are used, respectively, to modulate each curve in intensity and in duration. The nominal curve is defined by a set of key-points. Once it is loaded, interpolation for playing is made by a spline-based algorithm.

1) Example of a Quason Description: Let's see the following quason from the repository related to the SU of laughing: "laugh_begin_1." Each parameter is detailed in Figure 6.

On the left, the figure shows the definition of the structure of the quason. The first line corresponds to the name of the quason. Each of the following lines defines two parameters

[1]OSC: Open Sound Control is a standard protocol for communicating music devices on a network. It is an open alternative to MIDI.

[2]https://puredata.info/

and the normalized curve used. The parameters are used to modulate the curve in its *y*-axis in different ways for amplitude, frequency and timbre. Following the example, amplitude would be defined by the *cos_4.env* curve that is shown in Figure 5. This envelope is modulated to be between 0.7 (maximum) and 0.5 (minimum). The frequency of the quason will be centered at 750 Hz with a variation of 30%, which means between 525 Hz and 975 Hz. Timbre will be centered at 0.4 with a variation of 100%, that is, between 0 and 0.8.

The last line defines the duration of the quason in milliseconds, which is used to modulate the envelopes in time.

The right side of the Figure 5 is a graphical representation of a quason. Notice how the timbre is represented by the width of the pitch curve.

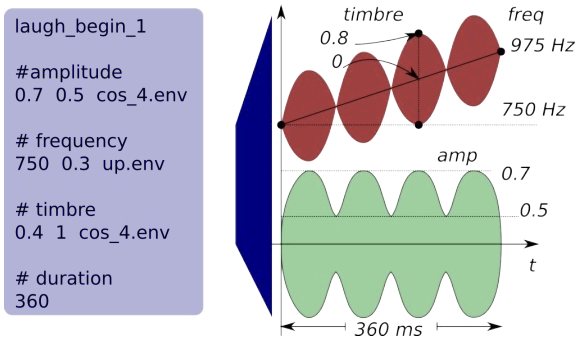

Fig. 6: Graphic representation of laugh_begin quason. The amplitude curve is cos_4, with values between 0.5 and 0.7. The frequency envelope goes "up-envelope" from 750 Hz to 750 + 30% (that is, 975 Hz). Timbre is another cos_4 envelope with values between 0 and 0.4 + 100% (that is, 0.8)

2) Example of an SU Description: Each SU is defined as a structure that includes its acoustic parameters and the set of quasons that form the SU. For instance, the SU *laughing* is defined as the following structure:

```
bpm 300          # beats per minute
laugh_begin_1 99 2 # quasons list
laugh_begin_2 79 2
laugh_continue_1 0 1 92 0.5 90 0.5 89
```

Listing 1: Example of an SU for expressing "laughing"

The first line defines the velocity of the SU by the bpm parameter (beats per minute). The following lines belong to each of the quasons that make up the SU. In the example, the SU is built from three quasons: *laugh_begin_1*, *laugh_begin_2* and *laugh_continue_1*.

For each quason, a set of notes is defined as pairs of values: MIDI[3] pitch and pulse duration. Each quason is played as a note. In the example, the laugh_begin_1 is played as a *D*#6, that in MIDI is number 99, and for two pulses, that is, as a half note. The quason *laugh_continue_1* is played as a semiquaver (0.5 of a pulse) after a crotchet of silence (0 MIDI pitch).

This list of notes defines both the melody and the harmony of the SU.

[3]Musical Instrument Digital Interface is a protocol designed for recording and playing back music

B. Quason Implementation

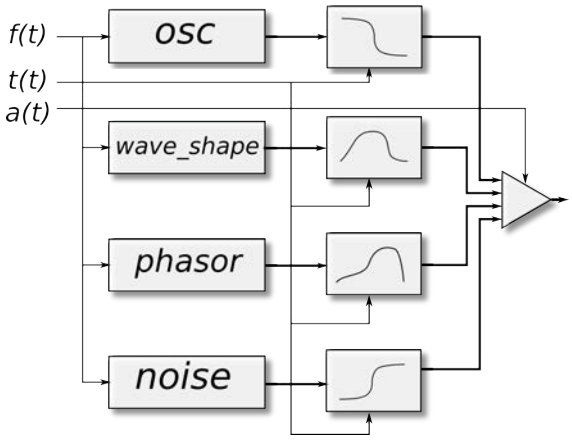

Fig. 8: Block diagram of the synthesis of quason timbre. The sound of a quason is a non-linear mixture of four sound sources: a simple oscillator, an oscillator of a custom wave shape, a square wave, and a pink noise. Each source takes as input the instantaneous frequency (*freq(t)*) of the quason, and generates a sound wave that is modulated by a non-linear gain function. This function depends on the instantaneous timbre (*tim(t)*). The mixed sound is modulated by the instantaneous amplitude (*amp(t)*)

Quasons are implemented using *puredata* [**?**], an open source visual programming language for music generation that enables researchers and developers to create musical applications graphically. *Puredata* can easily work over local and remote networks, which makes its integration easy in any existing architecture.

A quason is implemented as a set of *puredata* patches with objects that transform the acoustic quason features enumerated in Section III into sound, in real time. The amplitude envelope modulates the audio volume of the quason. The quason is modulated by the duration parameter, the pitch and the timbre envelopes. The timbre generation module, which is the core of a quason, will be explained in the following Section V-B1 in more detail.

1) The Generation of a Quason's Timbre: The sound of a quason is a weighted mixture of four different acoustic sources: a sinusoidal oscillator, an oscillator of a custom wave shape, a triangle wave and a pink noise source. Figure 8 shows how these four sources are mixed. The timbre generation subsystem receives three instantaneous variables: frequency (freq), timbre (tim) and amplitude (amp). Their values are obtained from the respective envelopes when the quason is played. So the pitch envelope will give a frequency variable.

2) Sound Generators: Four different sound generators have been considered in this paper: a sinusoidal oscillator, a pink noise generator, a phasor, and a custom wave generator. We give their details in the following.

The first generator is a *simple sinusoidal oscillator*, and represents the softest possible timbre, that is, a sound with just one frequency component. The *pink noise generator* produces noise with frequencies which are close to the quason's instantaneous frequency. The *phasor generator* creates a sawtooth

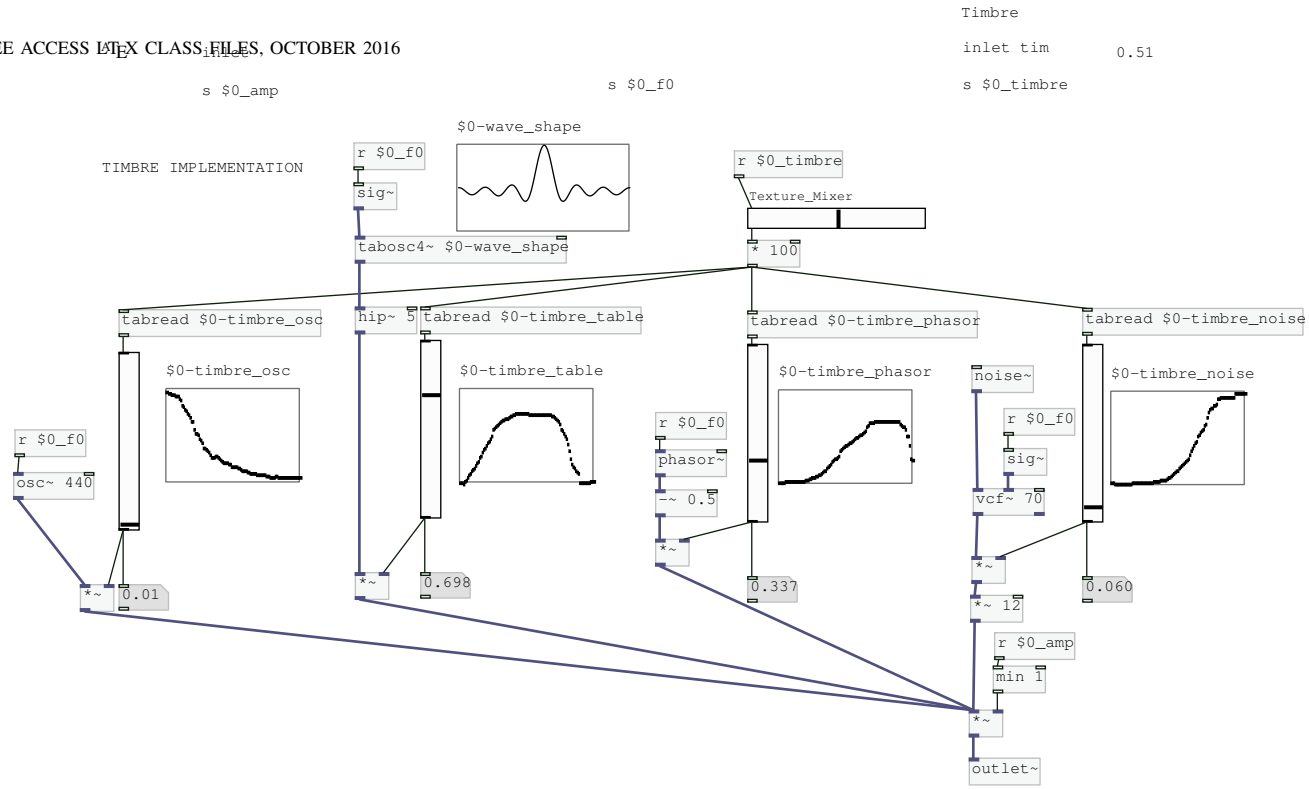

Fig. 7: Pure Data patch for the generation of a quason's sound timbre. The timbre is generated by four sound sources: an oscillator, a sawtooth generator, a cute sound source, and a pink noise generator. The sources are mixed by four empirical tables.

audio signal using the received instantaneous frequency. It sounds like a distorted sinusoidal oscillator and includes more harmonic components, so its sound is rougher. Its timbre could be considered to be between the sinusoidal oscillator and the pink noise. Lastly, the system allows generating an oscillator with any *custom wave shape*. This generator allows using a custom timbre. It is possible to load a wave shape taken from samples of an instrument (violin, piano, bell,..) or any other custom sound.

The signal of each of the four generators is modulated by a specific gain curve. Each gain depends on the instantaneous *timbre* value (*t(t)* as shown in Figure 8). For instance, in the sinusoidal oscillator the gain goes from a maximum value when $timbre = 0$, to zero, when $timbre = 1$.

The values of these gain curves were chosen empirically with the collaboration and advice of an expert musician. The main criteria used is that the final quason's timbre be an efficient and weighted combination that covers the essential necessities of an "abstract" granular sound going from the softest sinusoidal tone when $timbre = 0$, to the sharpest pink noise sound, when $timbre = 1$.

Figure 7 shows a partial *puredata* patch responsible for generating quason's timbre, so it implements the system presented in Figure 8. The patch receives three main parameters: the instantaneous amplitude (in normalized units), the fundamental frequency (in Hertz), and the timbre parameter (in normalized units). The timbre parameter is used to mix the four sound sources. The four different sound generators are implemented by four patches: *osc* ~, *tabosc* ~, *phasor* ~, and *noise* ~.

Patch *tabosc* ~ reads the wave shape to be synthesized from the *t_wave_shape* table. In the example shown in Figure 7, this shape is a harmonic combination of seven sinusoidal components, as expressed in Equation (1).

$$wave(t) = \sum_{n=0}^{6} (-1)^n 0.2 \cos(n f_0 t) \tag{1}$$

This wave is shown in the *t_wave_shape* graph. Its timbre is softer than a sawtooth but sharper than an oscillator, so it generates a "cute" sound.

Figure 7 shows how the four sound generators are mixed to create a sound with a specific timbre. The weighted mixing is made using four tables corresponding to the four gain curves: *t_timbre_osc*, *t_timbre_table*, *t_timbre_phasor* and *t_timbre_noise*.

3) Timbre Dimension as a Scalar Parameter: In the presented model, the quason's timbre is finally implemented as a normalized scalar variable. The system implemented in puredata allows modifying the timbre in real time. As explained above, the evolution of the quason's timbre is defined by the timbre envelope.

To depict how the timbre variable modifies the sound of a quason, Figure 9 shows the spectrum of a quason covering the overall range of the timbre variable, from 0 to 1 in six seconds. The spectrum swap begins with a sinusoidal signal at 440 Hz, with a narrow bandwidth, and continues adding more harmonic components and enhancing the bandwidth and the "color" of the sound. Finally, we have a pink noise centered on the fundamental frequency.

This spectrum is the result of how the four sound sources are mixed by their gain curves, as shown in Figure 8.

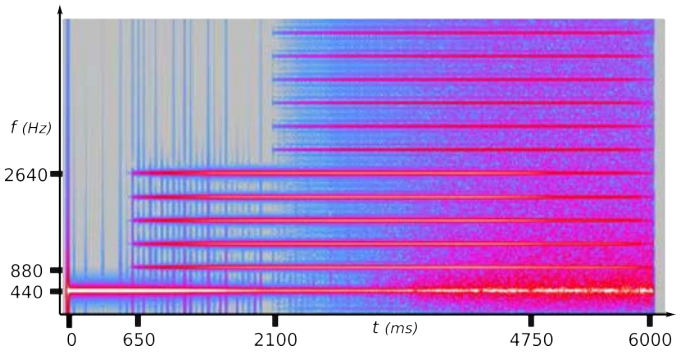

Fig. 9: Timbre swap spectrum from 0 to 1 uniform timbre values, 440 Hz fundamental frequency.

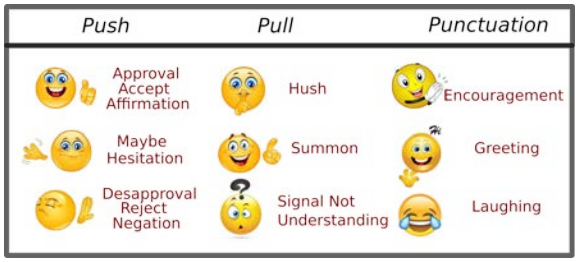

Fig. 11: Examples of SUs classified into three groups according to the pragmatic effect that the NVS tries to induce in the user: *Push* for giving information, *Pull* for asking for an action, and *Point* for pointing the sequence of the interaction.

C. An example of a Quason

The acoustic features of a set of sounds have been analyzed with the aim of representing such sounds in terms of the acoustic features of a quason.

Figure 10 presents one example of how one robotic sound is translated to quason acoustic features. The signal spectrum of the original sound is taken to create the frequency and timbre envelopes. The amplitude envelope of the quason is taken from the envelope of the signal. The figure also shows two tables with the key-points of frequency, timbre and amplitude used for creating the quason.

VI. Examples of Sonic Utterances Generated for the SES

To evaluate how the communicative expressions are understood by a non-expert user, we have implemented a set of NVS examples generated by the quason-SU formulation.

The concrete generated NVS are classified into three different communicative categories, depending on their intention of giving information to the user (the Push category), extracting information from the user (Pull category) or focusing on the flow of interaction (Punctuation category). Figure 11 shows nine natural communicative intentions that play a very important role in natural interactions.

The first column represents three communicative expressions that give or *push* some information, or just answer a question to the user: *affirmation*, *hesitation* and *negation*.

The second column represents three communicative expressions whose aim is to arouse, provoke, or *pull* a specific behaviour in the user: *hush* to keep silence, *summon* for asking that they move closer, and *signal non-understanding* or just *question* for repeating something that hasn't been understood.

Lastly, the third column represents three communicative expressions that are involved in how the interaction is sequenced or specified. In [**?**] it is called the "*punctuation* of sequences" in communication: *greeting*, which opens a new conversation, *encouragement* for expressing that the user is listening and supports what the speaker is saying, and *laughing* could explicitly value something as funny or could just strengthen the engagement in the interaction.

At the moment, we have not established a formal model for expressing each of these expressions. The sounds used for our test have been composed by the researchers with the help of a professional in music and sound art composition. The criteria used for such composition was based on previous studies in NVS systems, such as [**?**] or [**?**], and inspired by science fiction examples such as RD2D and Wall-E.

For each of these communicative expression, the sounds were composed and divided into three intensity levels of expressivity: low, normal and high. For instance, laughing is expressed in ways that range from an incipient smile to a loud guffaw.

The experiment was intended to test two aspects of the communicative efficiency of the SES. First, we tested the efficiency in the generation of NVS for very different communicative expressions. Moreover, we also tested whether the SES is able to express each communicative expression in different levels of intensity by just changing the acoustic parameters values that describe each sound. Note that some of the expressions would be implemented by a complex SU, others just by a simple quason.

A. Expression of Approval

This communicative expression includes any type of positive response, such as approval, acceptance, affirmation, agreement, etc. The results of [**?**] show that sounds for expressing agreement are more intense if they are shorter and do not change pitch. Moreover, we use the major mode, which better expresses a positive mood than does the minor mode. The rhythm pattern is chosen to express completion, conclusion, or resolution.

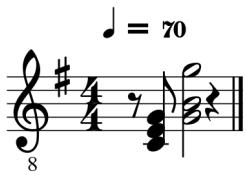

Fig. 12: Score for affirmation or approval at low intensity. Major mode in harmony and a rhythm pattern that expresses resolution.

The quason implemented for each note is an ADSR-based tone as follows:

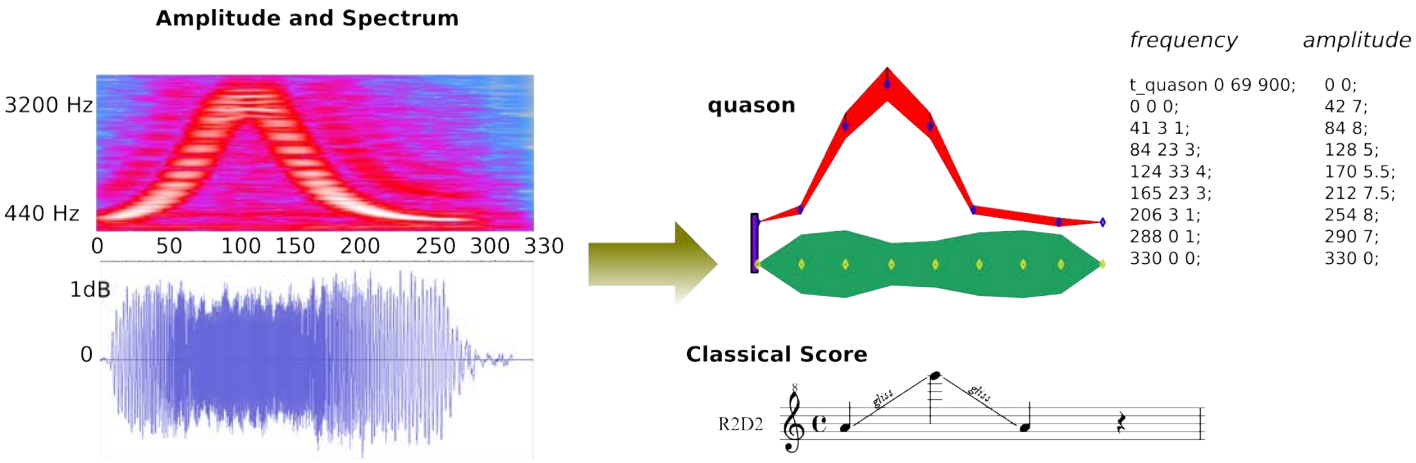

Fig. 10: Analysis of R2D2's beep and its representation as a quason. The quason is defined as the key-points of frequency, timbre parameter, and volume amplitude along time (in ms)

```
note_adsr           # quason name
0.5 0.5 adsr.env    # amplitude
440 0 const_1.env   # pitch
0.2 0 const_1.env   # timbre
```

Listing 2: quason structure for each note used in the Affirmation SU

Medium and high intensities for expressing approval are implemented keeping the same rhythm and harmonic pattern, but changing the tempo to ♩= 120 and 280, and lowing the key from G to E and C, respectively.

B. Expression of Hesitation

In essence, hesitation can be expressed by a constant pitch of a short sound. As shown in [**?**], the intensity of hesitation grows with the duration of the sound. The three levels of hesitation intensity have been composed as the following single quasons:

```
hesitation_0    # LOW level
1 0.5 adsr.env # amplitude
210 0.01 up_const.env # freq
0.01 0.2 cos_1.env # timbre
500     # durtion

hesitation_1    # NORMAL level
1 0.5 adsr.env # amplitude
220 0.0 up_const.env # freq
0.1 1 down.env # timbre
1000    # durtion

hesitation_2    # HIGH level
1 0.5 adsr.env # amplitude
230 0.01 down.env # freq
0.3 0.2 cos_1.env # timbre
2000    # durtion
```

Listing 3: Hesitation implemented in three different intensity levels.

Timbre also increases with the level of intensity, which means that for an increasing intensity of hesitation, a rougher sound is used.

C. Expression of Disapproval

Disapproval, denial, or rejection have been composed by playing with dissonance, as suggested in [**?**], using the triad $C\# - G - A$ (MIDI notes, 49, 55 and 57, respectively). In this triad there is a line of descending minor seconds with a simple rhythm pattern. This is the SU used for expressing disapproval at a high intensity level:

```
bpm 60
note_tremolo 49 5
note_tremolo 55 5
note_tremolo 57 5
adsr_up 0 0.5 72 0.5 71 0.5 70 0.5 69 3
note_tremolo 0 0.5 0 0.5 0 0.5 0 0.5 67 2
```

Listing 4: Disapproval SU in a high intensity

Normal and low levels are implemented by increasing the rhythm to ♩= 90 and 220, and shortening the duration of the SU. The *note_tremolo* quason is implemented as follows:

```
note_tremolo
0.2 1 tremolo_6.env # amplitude
440 0 const_1.env   # frequency
0.3 0 const_1.env   # timbre
```

Listing 5: quason "note_tremolo" implementation.

This quason uses a "tremolo_6.env" envelope for modulating the amplitude of its sound. This envelope is a sinusoidal curve with 6 as its maximum, so the sound amplitude *trembles* six times along its duration.

D. Expression of Hush

Hush is the sound for demanding silence: "sshhh." So an *ad hoc* quason with a high level of a constant timbre level has been used.

```
hush_0 # LOW level
0.1 0.2 up_down.env # amplitude
2300 0.01 up.env    # frequency
1 1 const_1.env     # timbre
600             # duration (ms)

hush_1 # NORMAL level
0.2 0.6 up_down.env # amplitude
2400 0.01 up.env    # frequency
```

```
1 1 const_1.env      # timbre
1000                 # duration (ms)

hush_2 # HIGH level
0.2 0.8 up_down.env # amplitude
2450 0.01 up.env # frequency
1 1 const_1.env # timbre
1200                 # duration (ms)
```

Listing 6: quasons for expressing ”hush” in three levels of intensity.

A high timbre value means that the signal is closer to pink noise, which best represents the “sh” sound. Intensity increases with volume and duration.

E. The Expression of Summon

To summon is to call for the presence of somebody. For instance, a whistle is normally used for calling a dog. We implemented each intensity by increasing the velocity and the pitch of a series of quasons that sound like a whistle.

```
bpm 90 # LOW level
summon 110 0.5 0 0.5 114 0.5

bpm 160 # NORMAL level
summon 112 0.5 0 0.5 116 0.5

bpm 220 # HIGH level
summon 116 0.5 0 0.5 120 0.5
```

Listing 7: SU for expressing summon in three different intensities.

The structure of summon quason is defined as follows:

```
summon
1 0.5 cos_1.env # amplitude
600 0.5 down_up.env # frequency
0.2 0.5 up_down.env # timbre
```

Listing 8: quason for expressing summon. It sounds like a whistle.

F. Expression of Lack of Understanding or Question

The expression of a lack of understanding is implemented as a short growing frequency.

```
adsr_up # quason name
0.5 1 adsr.env  # amplitude
440 0.01 up.env # frequency
0.5 0.1 down.env # timbre
```

Listing 9: quason for expressing Signal Not-Understanding.

The intensity is changed by means of the fundamental frequency: 66 (F4#), 69 (A4) and 73 (C5#), and the bpm from quicker to slower: 220, 120, and 90, as is shown below.

```
bpm 220 # LOW level
adsr_up 66 2

bpm 120 # NORMAL level
adsr_up 69 2

bpm 90 # HIGH level
adsr_up 73 2
```

Listing 10: SU for expressing Signal Not-Understanding in three different intensities.

G. Expression of Encouragement

For expressing encouragement, we composed a rhythm pattern in a bossanova groove[4] and a chromatic cadence with an ending growing sound, as is shown in the score represented in Figure 13.

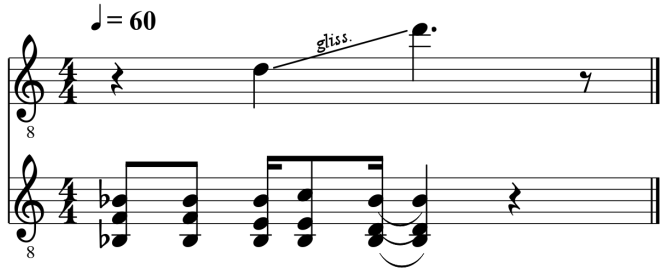

Fig. 13: Encouragement with high intensity

In this kind of rhythm, we use medium and low intensities, and we increase the tempo to ♩= 90 and 110, and finally we decrease the pitch by a semitone. Each of the four notes is implemented as an “ADSR_note” quason, defined in listing 2, the same as used for expressing approval.

H. Expression of Greeting

Greeting is an opening communicative act since an interaction often begins with a greeting. The composed sound is inspired by an onomatopoeia derived from how humans usually say “hello” as a sequence of two notes in a melodic resolution that represents opening.

```
bpm 90 # LOW level
note_tremolo 0 1 56 1
note_tremolo 0 1 64 1
adsr_up 56 0.75 0 1
note_up 0 0.25 68 1

bpm 200 # NORMAL level
note_tremolo 0 1 58 1 0 0.75 58 1
note_tremolo 0 1 66 1 0 0.75 66 1
adsr_up 58 0.75 0 1 0 1
note_up 70 1 0 0.5 70 1

bpm 55 # HIGH level
note_tremolo 0 1 60 1
note_tremolo 0 1 68 1
adsr_up 60 0.75 0 1
note_up 0 0.25 72 1
```

Listing 11: SU for expressing Greeting in three different intensity levels.

The low intensity greeting is represented as a musical score in Figure 14.

I. Expression of Laughing

Laughing has been composed by playing with the rhythmic pattern taken from a common guffaw. The SU shown in Section V-A2 corresponds to the sound that expresses laughing with a high level of intensity, see the musical score representation in Figure 15.

[4]A Brazilian rhythm that mixes samba and jazz.

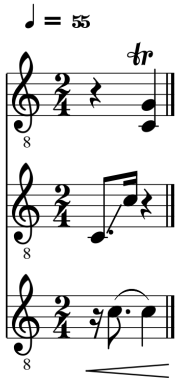

Fig. 14: SU score for expressing Greeting with high intensity

```
bpm 200   # LOW level
laugh_continue 80 0.5 78 0.44 77 0.40 75 0.3
  75 1

bpm 150   # NORMAL level
laugh_continue 82 0.5 80 0.44 79 0.40 77 0.3
  77 0.5 76 0.4 74 0.35 72 0.3 71 0.44 71 0.5

bpm 210   # HIGH level
laugh_begin_1 90 4
laugh_begin_2 70 2
laugh_continue 0 3 83 0.5 81 0.5
  80 0.5 78 0.5 77 0.5 75 0.5
  73 0.5 70 0.5 71 0.5 69 0.5 68 0.5
```

Listing 12: SU for expressing Laughing in three different intensity levels.

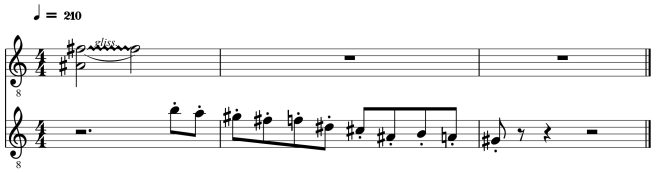

Fig. 15: Laughing in two parts: An opening chord and *marcato* descending quasons

First two fast quasons make the beginning of laughing as a chord, where the upper quason is a *glissando*[5], so the result opens the laughing. A chromatic scale of irregular *marcato*[6] quasons simulates the "ha" parts of a guffaw.

The SU uses three different quasons: two for the beginning chord, and one for the chromatic scale. Their structure has been implemented as follows:

```
laugh_begin_1    # quason name
0.7 0.5 cos_4.env # amplitude
3750 0.3 up.env   # frequency
0.4 1 cos_4.env   # timbre

laugh_begin_2    # quason name
0.5 1 up.env      # amplitude
1679 0.3 up.env   # frequency
0.2 1 up.env      # timbre

laugh_continue   # quason name
0.5 0.5 down.env    # amplitude
2318 0.05 up.env    # frequency
0.5 0.7 down.env    # timbre
```

Listing 13: Three quasons involved in the composition of SU for laughing.

[5]In music, a glissando is a glide from one pitch to another.

[6]This is a musical instruction indicating that a note, chord, or passage is to be played louder or more forcefully than the surrounding music.

VII. Evaluation of Our Sonic Expression System

The evaluation is focused on the identification of the SU by the users. The users tried to find the category of the played SU using three levels of intensity in the expression of each category.

A. The Robotic Platforms

In our experiments, we used the social robots of RoboticsLab, a research group at the University Carlos III of Madrid. They are: *Maggie* [**?**], *Mini* [**?**], and *Mbot* [**?**].

The Maggie robot, 1.40 m high, moves through the environment using a mobile base. The Mini robot is a "reduced version" of the Maggie robot, however Mini is not able to move in the environment. Mini has a height of about 55 cm and its external shell is covered by plush fabric. Mbot is a robot 1.05 m high and its shell is made of carbon fiber. This robot is a mobile robot and it is able to move around very quickly but still safely. All of them have interactive skills, and are equipped with broadcast-quality loudspeakers, microphones, and sound cards.

B. Experimental Setup

The users filled out a questionnaire about a set of 21 sounds corresponding to 9 different categories of communicative acts at 3 different levels: *low*, *normal* or *high*. It was possible to respond that the listened sound does not match any of the listed categories.

The sounds were played randomly to avoid order effects. As shown in Figure 16, each sound was played online as a different *YouTube* video and could be repeated several times if the user thought it necessary.

For testing the system, two main questions were set: how accurately the participants recognized each sound category, and whether the different intensities in each category affected the recognition rate. This was measured subjectively by a quiz of 9x4 multiple-choice questions, which corresponds to the nine sound categories (*affirmation, hesitation, negation, question, summon, hush, encouragement, greeting*) and the three intensity levels plus a "nothing" option[7].

Figure 16 also shows an optional *textbox* that opens the possibility for the user to add any new different communicative category label, in case he or she considers that the sound heard does not correspond to any of the initial set of nine categories.

[7]The on-line form can be checked on a dedicated web server http://soundsquestionnaire.web44.net/ or directly as an on-line file at http://goo.gl/forms/Kqh62wy9Gs.

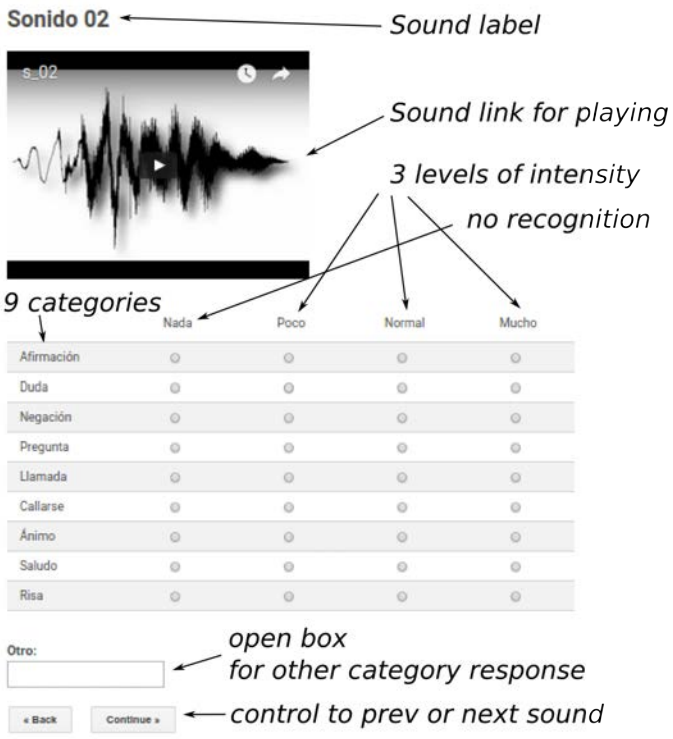

Fig. 16: Form used for the test. It represents the answers for 1 of the 27 SU. There is a multimedia link for hearing the SU followed by a matrix of responses. The user has to answer which communicative category best represents the heard sound, at three possible levels.

	aff	hes	den	q	sum	hsh	enc	grt	lgh
aff	**35**	3	3	10	5	3	18	21	2
hes	10	**26**	29	4	19	6	0	6	0
den	17	17	**26**	10	19	2	2	5	2
q	13	13	8	**31**	10	5	10	10	0
sum	5	8	2	7	**62**	7	0	9	0
hsh	2	12	12	2	12	**54**	0	6	0
enc	2	15	38	9	7	2	**14**	8	4
grt	5	9	25	19	9	2	1	**26**	4
lgh	2	10	16	9	2	3	5	5	**48**

TABLE II: Confusion matrix for recognition rates of **LOW** intensity sounds by communicative expression categories. Categories are: affirmation (aff), hesitation (hes), deny (den), question (q), summon (sum), hush (hsh), encouragement (enc), greeting (grt) and laughing (lgh).

	aff	hes	den	q	sum	hsh	enc	grt	lgh
aff	**30**	1	5	5	18	0	15	23	3
hes	20	**22**	18	14	14	6	2	4	0
den	0	21	**58**	6	4	4	4	0	3
q	13	16	8	**37**	7	0	6	8	5
sum	9	2	2	4	**47**	16	4	14	2
hsh	0	15	5	2	15	**59**	2	2	0
enc.	15	9	0	5	10	0	**30**	29	2
grt	8	10	24	11	26	0	5	**14**	2
lgh	2	7	21	3	3	3	5	5	**51**

TABLE III: Confusion matrix for recognition rates of **NORMAL** intensity sounds by communicative expression category. Categories are: affirmation (aff), hesitation (hes), deny (den), question (q), summon (sum), hush (hsh), encouragement (enc), greeting (grt) and laughing (lgh).

	aff	hes	den	q	sum	hsh	enc	grt	lgh
aff	**31**	0	6	2	12	3	23	21	2
hes	3	**16**	31	2	33	10	2	3	0
den	0	6	**86**	2	0	0	0	0	6

C. Results

A total of 51 participants participated in the study: 19 females ($ages = 38, SD = 10.1\%$) and 32 males ($ages = 37, SD = 10.0\%$) that completed an on-line questionnaire. All of them reported that they live in Spain.

The results are explained in three different tables according to the three levels of intensity of each SU. There is a great agreement between the subjects in their intentional interpretation of the SU. Some categories tend to be commonly confused, while others are more easily identified.

Tables II–IV show the results of the recognition rates for the nine categories of communicative expressions at the three levels of intensity. Each table has to be read as follows: "row sound category has been recognized as column communicative expression category." Table V shows the percentage rate of non-empty answers, that is, answers according to the three category levels: low, medium or high, and not to the option "nothing."

D. Discussion

According to the results presented in the previous section, we can summarize the following qualitative facts:

1) *Affirmation* is quite recognizable but it is confused with *Greeting* or *Encouragement*.
2) *Hesitation* is not easy to recognize, and it is confused with *Deny* and *Summon*, but never with *Laughing*.
3) *Deny* is the easiest category to be recognized, and its recognition rate increases with the level of intensity.
4) *Question* has been correctly recognized in the majority of cases. It is sometimes confused with any other category of the list, except *Deny*, *Hush* or *Laughing*.
5) *Summon* is quite easy to recognize, but it is sometimes confused with *Hush*.
6) *Hush* is also quite easy to recognize. The level of intention favors the recognition rate.
7) *Encouragement* at a low level of intensity is confused with *Deny*. At medium and high levels it is easier to recognize, though. Sometimes, it is also confused with *Greeting*.
8) *Greeting* is often confused with *Deny*, *Summon* or *Question*.
9) *Laughing* is easy to recognize regarding the rest of categories, but sometimes it is confused with *Deny*.

As mentioned above, not all the users answered all the questions. Table V shows the percentage of participation in each of the answers in the test. A lower level of participation indicates that the sound is not easily recognizable as a communicative expression category. Notice that in all cases, the percentage is above 90%.

On the other hand, some questions allowed open answers, i.e., users could respond by filling a textbox with their own appreciations, or adding a new category. These answers represent lower than 5% of the cases, but it might be interesting to mention it. There were three types of these open answers:

1) Synonymous, for instance, the user indicates "silence" in the Hush category (notice that an SU of this category is asking for silence); or "congratulations" in Affirmation;

Category	Level	Participation (%)
aff	LOW	92
	NORMAL	100
	HIGH	98
hes	LOW	94
	NORMAL	90
	HIGH	98
den	LOW	90
	NORMAL	82
	HIGH	92
q	LOW	96
	NORMAL	94
	HIGH	94
sum	LOW	98
	NORMAL	96
	HIGH	96
hsh	LOW	92
	NORMAL	100
	HIGH	92
enc	LOW	90
	NORMAL	96
	HIGH	96
grt	LOW	96
	NORMAL	94
	HIGH	100
lgh	LOW	96
	NORMAL	90
	HIGH	94

TABLE V: Percentage of non-empty answers in the form for the test. Each SU expresses a category at a level of intensity: low, normal or high. Categories are Affirmation (aff), Hesitation (hes), Deny (den), Question (q), Summon (sum), Hush (hsh), Encouragement (enc), Greeting (grt) and Laughing (lgh)

"indifference" or "tension" in Hesitation; "error" or "failure" in Negation; "mockery" in Laughing;

2) Emotional, for instance, "fear," "sadness," "weeping," and "deception" in Negation.
3) Other responses, for instance, indicating "farewell" for the low intensity Affirmation category, or "warning" for the normal level Question.

VIII. Conclusions and Future Research

We have formulated a theory of sound synthesis for communicative purposes in a systematic way.

A new Sonic Expression System (SES) has been designed and implemented. It includes the main acoustic features that give variety to the communicative possibilities of the system. All the systems in the state-of-the-art are pretty much simpler, being focused mainly on the expression of the prominent emotions, and are covered by our SES. This SES is based on normalized curves that are modulated in real-time for creating acoustic envelopes. These envelopes are gathered in a new concept in Non-Verbal Sounds generation: the quason. The system combines different quasons in a unique sonic phrase called a Sonic Utterance, that also has its own acoustic parameters/envelope curves.

Notice that the set of these curves could be easily increased. Also, these acoustic envelopes could be taken from any other source and not just from an initial domain. New envelopes are learned from perception skills. For instance, a robot could learn how to greet by perceiving and analyzing the sound of the user when he/she greets, and incorporate such envelopes in the *quason* domain of normalized curves.

In this paper, we covered the expression of nine common categories of communicative expressions by composing some *ad hoc* musical licks as SUs. We also presented and validated a system which allows having control over the degree of intensity of the expression of each communicative expression. Expressing some communicative intentions by NVS isolated from a context of reference is not an easy task. We found that some categories of communicative expressions are easily confused. For instance, "positive" categories such as Agreement, Encouragement and Greeting are sometimes confused with each other. But there are several categories that are very distinguishable from others, such as Deny, Laughing, Question (or Signal Not-Understanding), Summon or Hush. The results presented here should be improved if the SES is integrated in a complete multimodal robot which can use other cues than the sonic mode: gestures, semantic cues and dialog management, that help to set a coherent context of interaction.

In this paper, each sonic utterance is created by the developer with the main aim of demonstrating that the quason based model is able to express life-like communicative messages through the so-called sonic modality. Algorithm composition should allow generating such sounds automatically from a set of communicative intention features.

The system is conceived to be used in interaction scenes. The SES was built using *puredata* and allows real-time sound synthesis. However, the perception and planning skills have to be developed to be linked to the Continuous Composer, in the SES, for achieving the goals explained in I-A about natural interaction, and the interaction efficiency by means of message modulation and synchronization.

This paper could serve as a basis for additional experiments in non-verbal sound generation for expressing communicative intentions, involving more and different versions of SU in the expression of the nine categories studied here and/or others.

Acknowledgment

The research leading to these results has received funding from the projects: Development of social robots to help seniors with cognitive impairment (ROBSEN), funded by the Ministerio de Economia y Competitividad; MOnarCH, funded by the European Commission; and RoboCity2030-III-CM, funded by Comunidad de Madrid and cofunded by Structural Funds of the EU.

Javier Fernández de Gorostiza Luengo He works as assistant professor of the Systems Engineering and Automation Department at the Carlos III University of Madrid. He received the MSc Degree in Physics in 1999. In 2002 he also received another MSc Degree in Electronics Engineering. In 2010, he received the PhD degree in Robotics. His main research interests range from Human Communication and Interaction Models to Human–Robot Interaction. These include dialog systems, non-verbal interaction, multimodal interaction, etc. In the RoboticsLab, he leads two research topics: End-user Programming of Social Robots, and multimodal Human–Robot Interaction.

Fernando Alonso Martín He works in research and as a professor in the Systems Engineering and Automation Department at the Carlos III University of Madrid. His research fields are personal robots, human–robot interaction, dialogues, and other related issues. Currently he is mainly involved in two projects: RobAlz is a project born from the collaboration between the Spanish Alzheimer Foundation and the RoboticsLab. The aim is to develop robots that assist in the daily tasks of caregivers for Alzheimer's sufferers. MOnarCH (Multi-Robot Cognitive Systems Operating in Hospitals) is a European Union FP7 project that aims at the development of a network of heterogeneous robots and sensors in the pediatric area of an oncological hospital.

Alvaro Castro González He works as an assistant professor of the Systems Engineering and Automation Department at the Carlos III University of Madrid. His research fields are personal robots, human–robot interaction, decision making systems, emotions and motivations of robots and other related issues. Currently he is mainly involved in two projects: RobAlz is a project born from the collaboration between the Spanish Alzheimer Foundation and the RoboticsLab. The aim is to develop robots that assist in the daily tasks of caregivers for Alzheimer's sufferers. MOnarCH (Multi-Robot Cognitive Systems Operating in Hospitals) is a European Union FP7 project that aims at the development of a network of heterogeneous robots and sensors in the pediatric area of an oncological hospital.

Miguel Ángel Salichs He is a full professor of the Systems Engineering and Automation Department at the Carlos III University of Madrid. He received the Electrical Engineering and Ph.D. degrees from the Polytechnic University of Madrid. His research interests include autonomous social robots, multimodal human–robot interaction, mind models, and cognitive architectures. He was member of the Policy Committee of the International Federation of Automatic Control (IFAC), chairman of the Technical Committee on Intelligent Autonomous Vehicles of IFAC, the responsible of the Spanish National Research Program on Industrial Design and Production, the President of the Spanish Society of Automation and Control (CEA) and the Spanish representative to the European Robotics Research Network (EURON). He is the coordinator of the Secretariat of the Spanish Robotics Technology Platform (HispaRob). He is an Associate Editor of the International Journal of Social Robotics.