This work has been supported by the Spanish Government through the project called `Peer to Peer Robot-Human Interaction" (R2H), of MEC (Ministry of Science and Education), and the project “A new approach to social robotics" (AROS), of MICINN (Ministry of Science and Innovation). The research leading to these results has received funding from the RoboCity2030-II-CM project (S2009/DPI-1559), funded by Programas de Actividades I+D en la Comunidad de Madrid and cofunded by Structural Funds of the EU, developed by the research team RoboticsLab at the University Carlos III of Madrid.

Integration of a voice recognition system in a social robot

F. Alonso-Martín, Miguel A. Salichs

[1] University Carlos III of Madrid, System Engineering and Automation Department, 28911 Leganés (Madrid), Spain
falonso@ing.uc3m.es, salichs@ing.uc3m.es

Abstract — **Human-Robot Interaction (HRI)[1] is one of the main fields in the study and research of robotics. Within this field, dialog systems and interaction by voice play a very important role. When speaking about human-robot natural dialog we assume that the robot has the capability to accurately recognize the utterance what the human wants to transmit verbally and even its semantic meaning, but this is not always achieved. In this paper we describe the steps and requirements that we went through in order to endow the personal social robot Maggie, developed in the University Carlos III of Madrid, with the capability of understanding the natural language spoken by any human. We have analyzed the different possibilities offered by current software/hardware alternatives by testing them in real environments. We have obtained accurate data related to the speech recognition capabilities in different environments, using the most modern audio acquisition systems and analyzing not so typical parameters as user age, sex, intonation, volume and language. Finally we propose a new model to classify recognition results as accepted and rejected, based in a second ASR opinion. This new approach takes into account the pre-calculated success rate in noise intervals for each recognition framework decreasing false positives and false negatives rate.**

Keywords— **robot audition, automatic speech recognition, ASR, voice recognition, speech recognition, Maggie, personal robot, social robot, human-robot interaction, human-computer interaction, dialog, microphone system, audition system, natural language understanding, natural language processing, computing confidence score.**

I. INTRODUCTION

As human beings, we have five basic senses –sight, touch, hearing, smell and taste– that allow us to sense the surrounding environment and create our own mental picture of the world. What may result paradoxical is the fact that, despite of the fact that voice and therefore hearing is the most common way of communication in between people, the interaction with electronic systems it is usually done by means of written symbols (Jansen and T Belpaeme 2006), therefore related to sight. In the field of robotics, the interaction that takes place in between robots and humans is also studied: Human-Robot Interaction (HRI).

Research in social robots is a field that is clearly expanding(Hegel et al. 2009) , in which (multi-modal) Human-Robot Interaction plays a main role. In this field, there are several open lines of research: interaction via natural language, recognition of human gestures, built-in behaviors, and cognitive robotics (Sofge et al.). Technological advances in other areas benefit this research, in a way that it is increasingly possible to obtain an interaction quite similar to the one that is established between actual human beings. Among the technological advances of interest for this paper are the ones related to voice processing technologies.

These, namely, voice processing technologies have been developed to improve computer accessibility and also to interact with remotely located voice response systems. In our case, they are used to facilitate the communication with robots by sending and receiving oral information. In order to accomplish this objective, a minimum of three basic technologies are needed: technologies that enable written information to be transformed into spoken words *–text-to-speech conversion–*, technologies that allow a computer system to translate the spoken audio into robot action petitions *–voice recognition–*, and technologies that allow spoken interaction between a person and a service *–a dialog system* (Gorostiza et al. 2006) (Wallis 2010) (Shuyin et al. 2004)(Lopes and Tony Belpaeme 2008)– which is actually based on the two previously cited technologies.

One of the HRI areas of research focuses on trying to implement natural and agile dialogs in between the user and the robot, in a way that makes it possible to extract the relevant information from the conversational context in which they take place. There are different techniques and approximations to reach this objective. The most common and well known, especially in the area of telephone applications, is that de-

[1] Human-Robot Interaction (HRI) can be defined as the study of humans, robots, and the ways they influence each other.

fined by the VoiceXML[2] standard (Niklfeld and Finan 2001)(Nyberg et al. 2002)(Bennett et al. 2002) in any of its different implementations (mixed initiative, dynamical adaptative stategy…). Such Spoken Dialog Systems (SDS) are based in a finite state machine and aims to fill information gaps. There are also other alternatives, like, for example, those based on states using Partially Observable Markov Deccision Process (POMDP) algorithms (Roy, Pineau, and Thrun 1998), they are more flexible, maintain a probability distribution over the set of possible states (parallel dialog state hypotheses) and aims the improve the overall dialog accuracy.

ASR can be understood as the process of capturing and converting an acoustic signal, from a microphone input, into a string of written words, using a computer. ASR based technologies are classified into two basic types: "*speaker-dependent systems*" where the system is trained to recognize the voice of one specific user and will only recognize the voice of this specific user (here, recognition is open, in the sense that any sentence is possible and allowed; these systems are commonly known as dictating machines), and "*speaker-independent systems*". This kind of system is capable of recognizing sentences that meet specific sets of grammatical rules, spoken by any user without necessarily having previously trained with the system. In the field of HRI interests are mainly centered in this second type of system.

Throughout literature, for any automatic speech recognition system, it is common to describe different linguistic levels of abstraction: lexical, syntactic, semantic and pragmatic levels (Llisterri et al. 2003). The scope of these levels of abstraction reaches from phoneme joining for word construction, rules for word positioning within the context of a sentence, to the semantic meaning of the sentence in the conversational context and its relation with the general discourse. In the work here presented we focus on the first levels: lexical, syntactic and semantic, leaving the pragmatic level (related to the general discourse) aside.

This set of literature is focused on two important topics: "Robot Audition"(Valin, Rouat, and Michaud)(Tamai et al. 2005)(Hiroshi G. Okuno 2007)(Kazuhiro Nakadai et al. 2000) (as a subset of HRI) and "Natural Language Understanding" (Chan 1995)(Yu et al. 2009)(Junlan 2010)(J Li and Wang 1993) (as a subset of Natural Language Processing), but usually no information is provided regarding the software used to implement the automatic speech recognition part, or about the hardware used for audio acquisition. In the cases where this information is actually provided (Ishi et al. 2006)(H Kim and Choi 2007)(Kibria and Hellström 2007)(J Huang, Ohnishi, and Sugie 1997), we have detected certain inadequacies: they are mainly oriented towards usage on mobile robots, and therefore the social aspect is of secondary interest. Another issue is that they have been designed to deal only with English. Consequently, it is difficult to obtain and extend conclusions to usage with other languages. Finally, the most important deficiency, in this studies or surveys, is that no complete, comparative and precise study on the quality of voice-recognition capabilities using the different software and hardware options that are currently available, and testing over a variety of environments (including parameters as age, sex, intonation and language), is performed. On the other hand, we do not want to a present only a survey about microphones and recognition frameworks, but also we want to show the steps and features that are involved in the whole process of endow a robot the ability to recognize audio in a social context.

Our objective in this paper is to provide the guidelines and a general approach regarding which aspects and technologies should be taken into account when it comes to endow a robot with the capability to understand spoken natural language. Our objective has not been to develop the algorithms and basic functional components from scratch. It has instead been to integrate current technologies into our control architecture in order to get the highest possible quality in voice enabled HRI.

The work here described is only a portion of a much greater set that together forms a complete, modular and voice-interacted system where the most important modules are: automatic speech recognition (ASR skill), automatic voice synthesis (emotional TTS skill), localization of the audio source (Voice Tracking Skill), speaker identification (SI) and personalized dialog management (Dialog Skill).

[2] It should be noticed that VoiceXML is the W3C specified XML standard for interactive voice dialogs between a human and a computer. As opposed to the HTML standard that uses the screen and the mouse as the basic interface, VoiceXML instead uses speech, based on two mechanisms: Automatic Speech Recognition (ASR) and Text-To-Speech (TTS) conversion.

Some other similar and relevant works being developed are: the HARK open-source library (Kazuhiro Nakadai et al. 2008), which includes audio source localization modules, ASR and multichannel recognition, and even though it lacks the functionality required for speaker identification and dialog management, it implements similar functionalities by means of some low level software algorithms; other recent studies leading to a new line of research, are those that combine visual information, obtained from lip reading, with that from the audio in order to improve the recognition of the global input (Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno 2009).

Currently, there are many robots enabled with automatic voice recognition capabilities. Based on this capability, they can be classified into two types: first, those called “chatbots” or “virtual robots”, which “are alive” only on a computer screen, but lack a physical body. Some examples of these are Vikia, Grace, Valerie and Robotceptionist. On the other hand, there are those endowed with a real body. Among them, Honda ASIMO (Sakagami et al. 2002), SIG2, Robovie (Mitsunaga et al. 2006), and HRP-2 (Takahashi et al. 2010), are of great relevance to HRI. All of them incorporate the HARK audition software system. IROBAA (H Kim and Choi 2007) is capable of localizing audio sources by fusion of audio and visual sensory information. JIJO-2 (Fry, Asoh, and Matsui 1998) has been designed to live with humans in domestic environments, as has Robovie. They have the capability of learning the names of certain objects and places by speech, having some sort of semantic memory. HERMES (Bischoff and Graefe 2004) is able of understanding natural language, and has been tested for a period of 6 months as a museum guide. In fact, all of these robots are able to understand spoken natural language, but only over a subset of English and/or Japanese.

It should be noticed that robustness and high performance are primary objectives in ASR, but until now we have scarcely mentioned another fundamental aspect associated with this type of interaction: the audio capturing system. Typically, there are three types of microphones used in robotics. The first type is the “headset”, or unidirectional microphone, which is capable of capturing sound in one direction and only a few centimeters away from the mouth of the user. This kind of solution is the most common in current robotics. The second type is the “omnidirectional” or ambient microphone. It is mostly used to capture ambient sound; therefore its usage in HRI is very limited.

The last kind is the “microphone array”. It consists in any number of microphones (typically between 3 and 8 microphones) operating in tandem, fixed in a solid structure. The microphone array’s main features are extracting voice input from ambient noise (with noise reduction incorporated), and locating the sound source within a range of 1 to 3 meters. These features are very interesting in HRI and have been studied in several recent robots (H Kim and Choi 2007)(TAM and AI, Yoko SASAKI, Satoshi KAGAMI)(Valin, Rouat, and Michaud 2004)(Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno 2009)(Tanaka et al. 2010). This modern audio collection system is starting to be used more and more in videogames stations (Chetty 2009), laptops, mobile phones, cars (Oh, Viswanathan, and Papamichalis 1992), etc. See Fig. 1 for basic examples on robots.

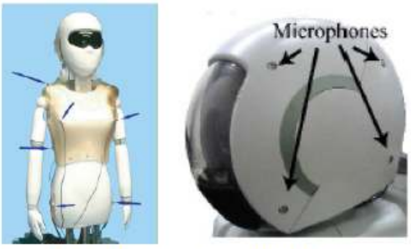

Fig. 1 SIG2 and ASIMO robots with built-in microphones

Robots should have hearing capabilities equivalent to ours to perform HRI in a social context, but in real environments there are many sources of noise. Many robot systems for social interaction avoid this problem by forcing the attendants of interaction to wear a headset microphone. For more natural interaction, a robot should listen to sounds with its own “ears” instead of making attendants use headset microphones (Breazeal 2003). For this purpose, in Maggie we are currently working on using a microphone array system, but we however consider it necessary to study the three previously mentioned types of microphones in several environments.

The imperfection of any speech recognizer reflects the reality that the state-of-art recognition systems still face problems in understanding spontaneous speech in noisy environments, hence one of the main challenges in the development of a robust ASR system is to deal with noisy input. A key step in addressing this noisy input is the computation of confidence (Lin and Weng 2010). For this reason, we finally propose a new model to classify recognition results as accepted and rejected, based in a second ASR opinion. This new approach takes into account the pre-calculated success rate in noise intervals for each recognition framework decreasing false positives and false negatives rate.

The following section of this article is a brief description on Maggie with details on the robot's hardware and software. Next will be a description of the technical requirements necessary for the system integration. In Section IV, we compare the most sophisticated commercially/open source available speech recognition packages based on the requirements described, and choose one of them. In Section V we explain how the ASR has been integrated within the robot's control architecture. We continue in Section VI testing the voice system in a test bed in real environments, with different systems for capturing sound. In Section VII is proposed a new experimental way to accept/reject utterance to comparing them with outputs of ASR engine. Finally, in Section VIII, the conclusions and future work is expressed.

II. WORK CONTEXT

A. The robot Maggie

The robot Maggie is a platform for studying HRI. The development of the robot is focused on finding new ways to adapt the potential that robotics has to provide new ways of working, learning and entertaining to human users.

An illustration of Maggie can be seen on Fig. 2.

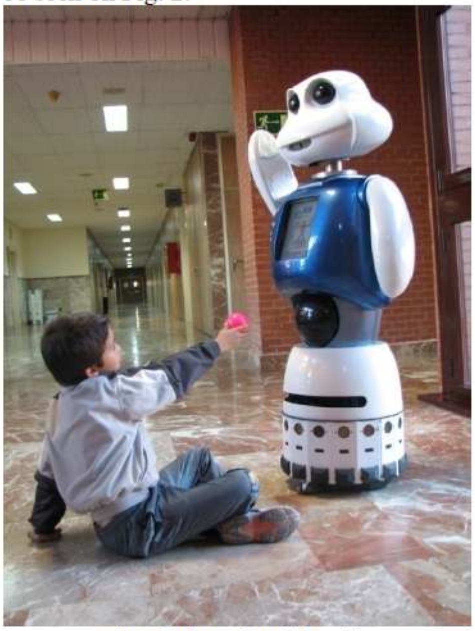

Fig. 2 The robot Maggie

HARDWARE

Maggie is designed as a 1.35 meters tall, girl-like doll. Its base is motorized by two actuated wheels and a caster wheel. The base is equipped with 12 bumpers, 12 infrared optical sensors and 12 ultrasound sensors. Above the base, a laser rangefinder (Sick LMS 200) has been added. The upper part of the robot incorporates the interaction modules. On top of the platform, there is a robot head with an attractive design. The head has two Degrees of Freedom (DoF), while each arm has one DoF.

Maggie is controlled by a main computer hidden inside her body. The software architecture of the robot lies inside this computer. For image acquisition, the robot has a camera located in the robot's mouth. The camera is a Logitech QuickCam Pro 9000. The robot has touch sensors on the surface of the body and a touch screen situated on the chest. Finally, inside the head, an RFID antenna is placed to identify objects.

SOFTWARE

The software architecture of the robot is the Automatic-Deliberative architecture (AD). AD is composed by two levels, the automatic level and the deliberative level. The automatic level is where the low-level control is performed: in the automatic level, the modules that provide communication and control of the sensors, motors and other hardware are located. At the deliberative level, reasoning and decision processes are placed.

The essential component of the AD architecture is the skill. A skill is an entity that is able to reason, process data or perform actions, and is able to communicate with other skills (similar to what occurs in the Hermes Skill-based system architecture (Bischoff and Graefe 2004)). A more detailed description of the AD architecture can be found throughout the authors' previous publications (R. Barber and Ma Salichs 2002)(R. Rivas, A. Corrales, R. Barber 2007).

III. REQUIREMENTS FOR VOICE SYSTEM

As said in the Introduction Section, the main goal pursued in this work is to show the steps we have taken to integrate an ASR system in a social robot and improve HRI with this voice system. This involves studying software and hardware technologies and solutions. In our control architecture, the ASR capability must be implemented and integrated as what we call a "Skill", and must allow any component of the control architecture to be able to use the voice recognition functions easily. In order to achieve this goal, we should define the requirements that are necessary to achieve a good, modern and powerful ASR system:

MINIMUM REQUIREMENTS

- Speaker independence: The system must recognize the natural language spoken regardless of whom is the person who is speaking, without need of prior training.

- Highly accurate speech recognition: Recognition results should be as accurate as possible. In an ideal case, the speech recognition system's accuracy should be similar to that of a normal person to understand what another person is saying.

- Support to the PC microphone: Usually the recognition systems are designed for telephone applications. However, we need to obtain the audio signal from the PC microphone input. The microphone must be continuously sending audio samples to the speech recognizer (streaming). Telephone applications send audio sample (complete) files to the recognizer.

- Operating system support: If the robot system architecture must necessarily run on Linux, we will need a Linux compatible ASR software.

- Change grammars in real time ("on the fly"): Speech recognizers are based on "templates", which indicate the valid rules and combinations of the audio input for the linguistic context, called grammars. It must be possible to change the grammar, add a new grammar or remove a previous grammar even when the recognizer has already been initialized.

- Speech detector: The system must be able to distinguish between voice and noise, so it must have a noise cancelling system, usually it is based in a noise threshold. An extreme scenario would be a case where the ambient noise was higher than the volume of the human voice, and the speech detector is able to distinguish between voice and noise. With speech detectors, it is not necessary to press any button to notify the robot when we start and finish talking; the robot can be constantly listening.

DESIRED REQUIREMENTS

- Support semantic grammars: Semantic grammars make it easier to extract the information that is relevant from complete sentences that have been recognized. Semantic grammars differ from normal grammars that include post-processing of the information recognized, this post-processing task is carry out by a scripting language built-in the grammar file, it allow achieve the semantic level. For better understanding of this concept, you can read about Natural Language Processing or Natural Language Understanding (Walker 1976)(Favre, Bohnet, and Hakkani-Tur 2010a)(Favre, Bohnet, and Hakkani-Tur 2010b)(Valverde-Albacete and Pardo 1996)(Lecouteux, Nocera, and Linares 2010)(K Kim, Jeong, and GG Lee 2007).

- Support standards: Several standards that help develop speech applications in a more simple and standardized way have been in defined speech technology. The most important formalism and standards in ASR technology are:

- SRGS: Speech Recognition Grammar Specification
- NLSML: Natural Language Semantics Markup Language
- SISR: Semantic Interpretation for Speech Recognition

- High efficiency: The possibility of use the recognition with a low computational power consumption allows the machine CPU to not be completely busy executing the speech recognition system. Moreover high efficiency allows a fast response by the recognizer engine, providing results within a few milliseconds, which is very important in Human-Robot Interaction.

- Multilanguage support: It has to be able to make language recognition in several languages and dialects. In our case, at least we need to recognize in Spanish, American English and British English languages.

- Speaker identification: This feature provides that the system is able to distinguish the speaker from among a group of potential users while the recognition utterance also is performed. It is an important feature that can be used in a enroll phase. VoiceXML 3 standard says: “The acoustic verification may compare speech samples to an existing model (kept in some, possibly external, repository) of that speaker's voice. A verification result returns a value indicating whether the acoustic and knowledge tests were accepted or rejected. Results for verification and results for recognition may be returned simultaneously”.[3]

- Acoustic model adaptation: Usually the recognition engine is trained in telephone environments, so the possibility of re-train the model for our real scenario using our own hardware to increase the accuracy is often desirable. It is necessary especially in array-microphone systems.

- Statistical Language models for dictation: Sometime we may want to use the recognition engine as a simple dictation tool, in technical words, without using restrictive grammars and without extracting semantic information. In this case, it is necessary to use language models to get high accuracy in dictation. Language models use dictionaries and sets of possible sentences, and can be based on bigrams or trigrams. In bigrams the probability of a word within a sentence is conditioned by the preceding word, while in the trigrams the probability of a word is conditioned by the two preceding words. With a huge language model the ASR can be done without a personal training phase for each speaker, instead in other cases is necessary a customized training phase with the recognition system and each user speaking some utterances and sentences for high success rates.

- Partials results in recognition phase: Sometimes it is very usual that the user is speaking for several seconds saying a long sentence. In this case is very convenient that the system can provide partial recognition results as soon as possible.

- Support tools: Tools to measure the efficiency of voice recognition, to help to write valid grammars, to generate logs or billing, to compile grammars, etc.

-Technical support: In the sense of the company that develops the voice recognition framework (the supplier) providing fast support to developers to make their work easier.

IV. OVERVIEW OF AVAILABLE ASR FRAMEWORKS

Once we have defined the requirements that our recognition ability must have, we needed to study what software solutions are best suited to such requirements.

[3] http://www.w3.org/TR/vxml30reqs/#funct-siv

Of all the systems available, both commercial and free, based on literature (Kibria and Hellström 2007) and our own experience, we have selected the best known and potentially most powerful ones to make a detailed study on each of them, comparing them in all of the aspects identified in the previous section.

Finally, the five systems under comparative study were the following:

- Verbio ASR v8[4].
- Nuance Recognizer V9[5].
- Nuance VoCon 3200[6].
- Loquendo ASR V7.7 (patch 25)[7].
- Sphinx IV[8] (XD Huang et al. 1991).

A test license for performing analysis in our installations was acquired for each one of these systems. The documentation of each one of the frameworks was fully analyzed, the programs were installed and configured to work in our environment, and a test suit was run to analyze the performance in each aspect of the proposed requirements.

The fact that obtaining all of the licenses, analyzing the large amounts of documentation, and setting up each specific system for running in a real working environment should be taken into account as non-trivial or easy task. However, these steps were necessary to be able to make a rigorous comparative analysis of them.

Most of the details of the survey we conducted have been summarized in Table 1. In this table, each column represents a different recognition framework, and each row represents the value taken for each of the conditions analyzed. Most of the evaluated characteristics were determined in an objective manner, as in determining whether a product includes a "speech detector" or not. However, other features, such as the usability of the product, are the result of our own subjective experience in using such tools.

	Verbio ASR	**Nuance Recognizer V9**	**Nuance VoCon**	**Loquendo ASR**	**Sphinx IV**
Developed in …	Spain	EEUU	EEUU	Italy	EEUU
Without training	Yes	Yes	Yes	Yes	No
Speaker-independent	Yes	Yes	Yes	Yes	Yes
Grammar-based	Yes	Yes	Yes	Yes	Yes
Statistical Language Model	Yees	Yes	No	Yes	Yes
Operating System Supports	Linux /Windows	Linux /Windows	Windows	Linux /Windows	Linux /Windows /Mac OS / Solaris
Embedded System support	Yes	No	Yes	Yes	Yes
Speech Detector	Yes	Yes	Yes	Yes	Yes
Microphone support	Yes	No	Yes	Yes (with addon)	Yes
Multilanguage	Yes	Yes	Yes	Yes	Yes
Phoneme based	Yes	Yes	Yes	Yes	Yes
Speaker identification	No (requires another SW)	No	No	Yes	No
Word spotting mode	Yes	Yes	No	No	No
Semantic models	No	No	Yes	Yes	No
New Words Learning	No	No	Yes	No	No
User Acoustic Model	No	Yes	No	Yes	Yes

[4] http://www.verbio.com/

[5] http://www.nuance.com/for-business/by-solution/contact-center-customer-care/cccc-solutions-services/recognizer/index.htm

[6] http://spain.nuance.com/vocon/

[7] http://www.loquendo.com/es/technology

[8] http://cmusphinx.sourceforge.net

Adaptation					
Usability	High	Low	Very High	Low	Medium
Examples	Normal	Normal	Very good	Poor	Hight
Additional resources	Poor	Normal	Very good	Good	Normal
Support	Yes	Yes	Yes	Yes	No
Easy to buy	Easy	Difficult	Difficult	Difficult	Very easy
Effectiveness/Accuracy	**Medium:** 91% (85%)	**High:** 99% (94%)	**High:** 99% (86%)	**High**: 99% (98%)[9]	**Medium**: 94% (87%)
Price	Low (lower than 1.000$).	High (about 5.000$)	Medium-Low (about 1.500$)	Medium (about 1.500$)	Free (0€)

Table 1 ASR Framework comparison table

Based on this survey and on needs for improve HRI, we decided to choose the Loquendo framework for integration within our control architecture. The choice is justified by the compliance of Loquendo with all the requirements outlined in the previous section. Another aspect that has additionally supported our decision is that it seamlessly integrates with the speech synthesis software, which is also developed by Loquendo, on which our speech synthesis ability (not presented in this article) is based. On the other hand the relation between accuracy/cost placed him as leader.

The remaining products were discarded due to the following important reasons:

- Nuance Recognizer V9: although it is a great framework with a very good success rate, it is designed for telephone applications and to date, it does not support directly audio input through the microphone and it is the most expensive framework.

- Nuance Vocon: other great Nuance product with a wonderful performance, but to date, there are not Linux version available yet.

- Verbio ASR V8: It is the lowest accuracy framework and no partials results are provided for it.

- Sphinx IV: a open-source development, with a great language and operative systems support. It is being developed by Carnegie Mellon Universtiy, Sun Microsystems, Hewlett Packard among others, but the major problem is that it is less accurate that Nuance and Loquendo.

V. INTEGRATION WITH THE ROBOT CONTROL ARCHITECTURE

Once we have chosen the product that best fits our needs, we must integrate it within the robot control architecture. As we have said, in the AD control architecture, we call any software component that provides a new capability to the robot “skill”. Therefore, we must integrate the voice recognition system as a skill in the architecture.

The skill we have developed is structured in three layers of abstraction and thus of complexity. Situated at the lowest level is the recognition engine, which is composed by the framework and the libraries that are provided by Loquendo. These libraries are written in C programming language, and an additional Java wrapper is provided. Above of the recognition libraries, which perform the actual recognition capabilities, we have written some basic functions, which we call “ASRPrimitives”. These primitives implement the most important functionalities such as setting grammars, starting and stopping voice recognition, setting the

[9] We have achieved these results in silent environments and noise environments severally in a grammar-based mode and using expensive high-quality directional headphones (Senheiser HSP-2, http://www.sennheiser.com). The ‘%’ is the Success Rate; not the confidence given for the ASR engine and nor the success rate that claims the sales staff).

format of the audio input and obtaining the results of speech recognition. Finally, over these primitives, we have built the recognition skill, "ASRSkill", which has the format of all of the skills of the control architecture. Any skill that wants to perform speech recognition delegates the task to the ASR speech recognition skill, as seen in Fig. 3.

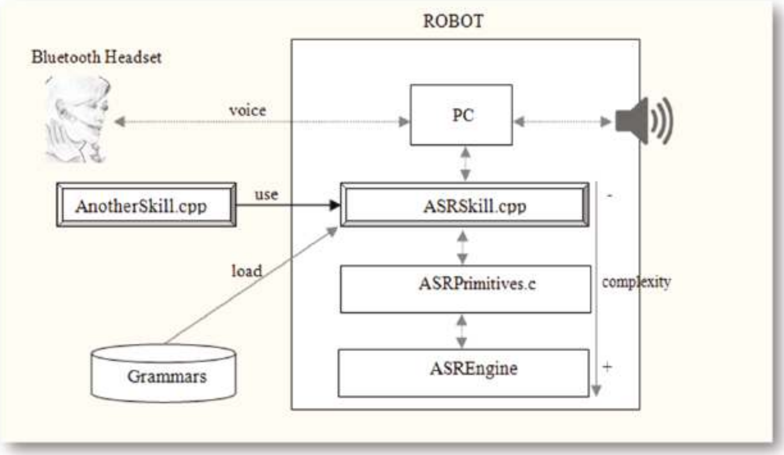

Fig. 3 ASRSkill layer structure

This layered structure corresponds to the three levels of language processing to understand a sentence in a dialogue context:

1º Level: Speech recognition
- Acoustic language models – words lists (built-in Loquendo)
- *What has the caller said?*

2º Level: Speech analysis
- Grammars – lexical meaning
- *What did the caller mean?*

3º Level: Understanding
- Discourse context – knowledge about domain of discourse.
- *What has the caller asked?*

The first level (speech recognition) matches with *ASREngine* provided by Loquendo. The second level (speech analysis) matches with *ASRPrimitives*. Finally, the 3rd level (understanding) matches with *ASRSkill*, that is a level above. *Finally,* over them, is placed the Dialog Managerar which controls *eTTSSkill*, *ASRSkill* and *identificationSkill*, and is based on the VoiceXML standard for control of the dialogue between the human and the robot in several languages. This modular voice system architecture can be seen in Fig. 4, and is a subset of all of our work.

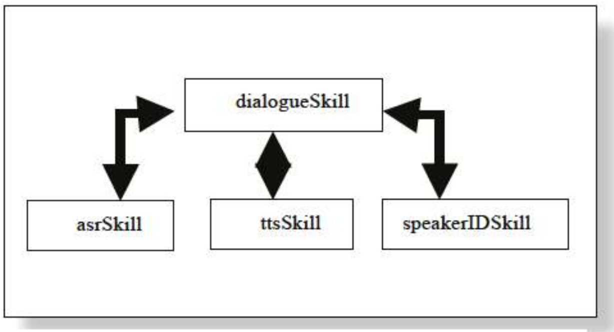

Fig. 4 AD voice structure

Certain typical technical aspects to be treated always require attention:

- In our case, the format of the audio samples must be ULAW (the other possible format is ALAW). They are the uncompressed audio de facto standard format in Unix sound[10][11].

- The grammars the skill can work with are localized in the laptop hard disk, and when the skill needs a specific grammar, it can be read and loaded into the main memory (RAM).

- The acceptation threshold must be tuned by the specific application and environment, but a good default value can be 0.50.

- The limit search space can be tuned too and it is used to give the possibility of controlling the recognition search space dimension to adjust the recognition accuracy versus recognition speed. The default value we use with our grammars is 500.

VI. ACCURACY EXPERIMENTS

Now that we have chosen a framework and we have integrated it into our robotic control architecture, we need to test the ASRSkill in real environments and with different hardware for capturing audio (different microphones). We have built several test scenarios in order to study the accuracy depending on several parameters: noise dependence, speaker volume, voice intonation, age, sex, and the importance of the type of microphone used.

Each test has been performed with different users saying different sentences, using the same grammar and without previous training with the ASRSkill. In the first test scenarios, we performed the audio acquisition with professional **unidirectional wireless headsets**[12]. The microphone is located a few centimeters away from the speaker's mouth and in the same direction. The acoustic signal is transported through the air and it reaches the robot through a receiver. The transmitter and receiver are show in see Fig. 5.

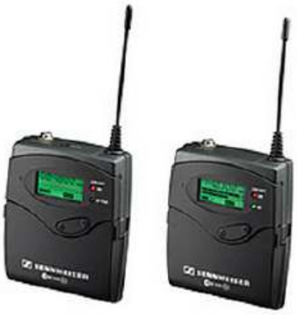

Fig. 5 Wireless microphone transmitter-receiver couple

[10] http://en.wikipedia.org/wiki/A-law

[11] http://en.wikipedia.org/wiki/M-law_algorithm

[12] The specific transmitter model used has been:
http://www.sennheiser.com/sennheiser/home_es.nsf/root/professional_wireless-microphone-systems_broadcast-eng-film_ew-100-series_021418
The specific headset models used have been:
http://www.sennheiser.com/sennheiser/home_es.nsf/root/professional_wireless-microphone-systems_headsets_headsets_009862 (500$) and
http://www.logitech.com/en-us/webcam-communications/internet-headsets-phones/devices/3621 (40$)

The users began by saying sentences which fit the grammar established to be used in the test. They said sentences continuously, without following any instruction. Meanwhile, we monitored the experiment and took notes of the results related to the speech recognition.

Once the user completed his turn of sentences, went the next user. Users knew the possible sentences they could say because they knew the test grammar. **Each test has been done on a group of 10 users and a total of 100 voice recognition.**

To estimate the accuracy of speech recognition, we have analyzed two parameters: the **success rate** and the **confidence value**. We consider the success rate is equal to the mean percentage of times that the recognizer is capable of matching the correct sentence. Similarly, we consider that the confidence value is the guarantee that the recognizer has performed correctly. If the confidence value is close to 1, the recognizer is almost certain that the recognized sentence matches with what the user has said. However, if the confidence value is very close to 0, this indicates that it is not confident in what has been recognized (possible mismatch with what the user has said).

With the confidence value of each recognition, and fixed a value threshold, we decided if clarification sub dialogues were required. We tried to avoid accepting the false recognitions and maintain the correct ones.

Success rate can be used to compare the accuracy of different recognition engines. As there is no value given by the recognition engine providers, therefore these values are extracted from the real tests performed, with multiple users, in different conditions and environments.

SENTENCES WITHOUT IMPORTANT NOISE BY HEADSET

In the first test scenario users communicated with the robot in a closed environment, in the laboratory, with no significant noise. This is, ranging from approximately 40 to 45 dB. This kind of noise is called stationary noise, and is produced by the robot, computers, fans… Moreover, it is very easy to predict and eliminate.

As results, we have obtained an average confidence value of **0.722**, where 0 is the minimum value and 1 a highest value. In **99 percent** of the cases, the sentences were accurately recognized (the sentence had to fit within the grammar established).

With these results we conclude that the speech recognizer's accuracy and with speaker independence is extremely high in silent environments. The confidence value is quite high, and the success rate is very close to 100. This means that recognition accuracy is almost complete with these conditions and with these microphones. Results match with the official Loquendo results (Paolo Baggia 2005)(Dalmasso et al.).

SENTENCES WITH BACKGROUND NOISE BY HEADSET

This test scenario is very similar to the one explained above, except that we added background noise to the test. We left a television set turned on and music in the background. This noise is emitted 7 meters away from the user in place with the microphone. The background noise is about 65-70 dB measured from the place where the human is.

The results are a **0.703** confidence value and a **98%** success rate. These values have decreased slightly compared with the above scenario, but the recognizer skill can still be considered very accurate. We can conclude that it is robust against background noise conditions using the same microphone configuration.

It is very important for the microphone to be located very close to the speaker's mouth, as we are using a unidirectional microphone that picks up sound from only a few centimeters away and in one direction. This kind of microphone is the most similar to those of mobile phones. The reader should additionally take into account the previously mentioned fact that recognition engines are trained for telephone voice applications.

SENTENCES WITH NOISE CLOSE TO THE HEADSET

In this test scenario we placed the source of noise very close to the user. We put a music speaker one meter away from the user place. The noise ranges between 70 and 75 dB.

The results that we have obtained are **0.673** in confidence value and **97%** rate success. These values are similar to those of the previous test scenarios. The main reasons for obtaining these high values are that unidirectional microphones only capture the user's voice and very little of the background noise, the Loquendo recognition engine has trained for noisy conditions (the acoustic model) and finally, the actual recognition engine can eliminate the stationary noise.

Therefore, we can conclude that in a noisy environment, even with very adverse noise conditions, the ASRSkill is very robust and accurate using the appropriate microphones. See Fig. 6 for a comparative line graph.

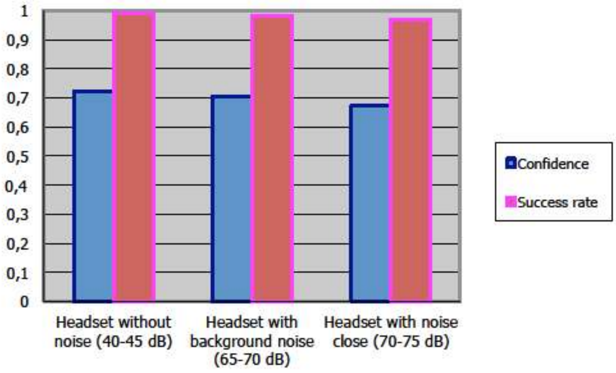

Fig. 6. Accuracy summary using headsets

SPEAKER VOLUME AND QUALITY OF SPEECH RECOGNITION BY HEADSET

In this test scenario we tried to analyze how the speaker volume affects the recognition accuracy. Speech recognitions were performed for a sentence said by each user 10 times, each in different volumes. We have measured the recognition accuracy and the speaker volume with a sound level meter at 3 cm from the speaker's mouth. We have obtained the following results:

Low volume (69 dB): **0.66** confidence value (100% success rate)
Medium volume (77 dB): **0.72** confidence value (100% success rate)
High volume (83 dB): **0.80** confidence value (100% success rate)
Very high volume (89 dB): **0.70** confidence value (100% success rate)

With these results we can conclude that, although the speaker volume affects the confidence value, the differences are not very large, and the success rate is practically the same in all cases. When pronunciation is clearer and the audio volume is the most intelligible possible (without distortion), precision is greater. We could say that whatever is more compressible for ourselves as a speech recognized is also more understandable for the automatic speech recognizer too. See Fig. 7 for a line graph representation of the results.

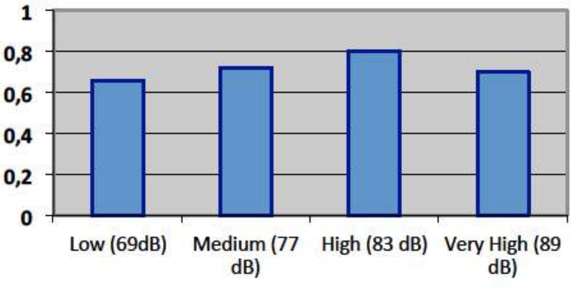

Fig. 7. Accuracy and speech volume

SPEAKER INTONATION AND SPEECH RECOGNITION ACCURACY BY HEADSET

Here, we tried to determine the relationship between the speaker intonation, saying the same sentence, and the accuracy of the speech recognizer. The same sentence is pronounced with different intonations: declarative and interrogative in a normal use of both (without exaggerating the intonation). The results were:

- Declarative sentences: 0.74 confidence value (97% success rate).
- Interrogative sentences: 0.71 confidence value (96% success rate).

With these results, we can say that the intonation is not a decisive factor that affects speech recognition. Although we usually use declarative sentences, other intonations are also properly recognized.

SPEAKER SEX AND QUALITY OF SPEECH RECOGNITION BY HEADSET

Other important test scenario is to see the sex of the speaker affects the accuracy of the speech recognizer. For this we evaluated speech recognition with different men and women, using the same grammar and in the same environment. The results are:

Women: **0.70** confidence value (99% success rate)
Men: **0.69** confidence value (98% success rate)

With these results, we can conclude that the recognizer is independent of the speaker sex. This is quite logical because the underlying neural networks were trained using the same proportion of men and women.

AGE OF THE SPEAKER AND QUALITY OF THE RECOGNITION BY HEADSET

In this case we tried to test the speech recognition with people of different ages. We divided the people in two groups: children between 5 to 12 years old, and adults from 13 years old to 70 years old.

The results are:

5-12 years: **0.522** confidence value (93% success rate)
13-70 years: **0.722** confidence value (99% success rate)

We can see that in the age group ranging from 5 to 12 years, the speech recognition is worse than in the other group. The confidence value and success rate is significantly lower. This is because children express less clearly and their voice is still less educated (more wean and shaky). However, success rate is still high, enough to interact with children. This problem has been described throughout literature (Ishi et al. 2006) and a solution for other systems, with worst confidence results, is using different recognizers with different acoustic model adaptations (one for children and another for adults).

MICROPHONE BUILT IN THE ROBOT (OMNIDIRECTIONAL MICROPHONE)

In all test scenarios described above we have used unidirectional wireless microphone headsets. In these cases, the speaker had to put the microphone very near the mouth. However, interaction without headsets is much more natural and comfortable (Breazeal 2003). There is a need for the robot to be provided with mechanisms to collect audio itself.

For this test we tested with an **omnidirectional microphone**[13] (or non-directional microphone) built in the robot. This kind of microphone (see Fig. 8) is able to obtain audio information from the environment in any direction. They are typically used to collect ambient sound, or to record music choirs. They are able to collect the audio from a few meters away with enough quality. The main problem is that they are much more sensitive to noise than unidirectional microphones because they are designed for a different purpose.

[13] http://en.wikipedia.org/wiki/Microphone#Omnidirectional

The advantage is that the user can talk to the robot without any additional device, achieving an interaction very similar to that that occurs between humans.

Fig. 8 Omnidirectional microphone MP33865

The test were performed using different distances from the robot (1, 2 and 3 or more meters), without important noise background (less than 50 dB). The results are:

1m: **0.42** confidence value (75% success rate).
2m: **0.31** confidence value (72% success rate).
3m: **0.25** confidence value (66% success rate).

The fact that the speech recognition's accuracy decreases with the distance to the microphone can be appreciated. Results are worse than using unidirectional microphones, but depending on the application, it might be sufficient.

We have not been able to obtain results in an environment with significant ambient noise (65-75 dB), because the speech recognition cannot differentiate between noise and speech. The problem is that the noise and the human voice arrive to the speech detector at similar volumes. The noise cancelling system of the ASRSkill eliminates both, the noise and the speech samples. Additionally, this kind of microphones lack noise cancellation systems (because they are designed to receive the ambient sound).

Another approach to improve recognition results is to train the speech recognizer for this particular acoustic model. It is important to remember that the speech recognizer acoustic model is designed and trained for telephone applications with unidirectional mobile microphones.

To sum up, this kind of microphones (non directional) eliminate the need of external devices for communication and provide a more natural interaction between human and robot. However, in environments with an important background noise, they are a poor choice for HRI.

MULTI-ARRAY MICROPHONE BUILT-IN THE ROBOT

A microphone array is any number of microphones operating in tandem. Their main applications are for extracting voice input from ambient noise and locating the sound source (the angle from which it is originated). These features are very interesting in HRI and have studied using recently using several robots (H Kim and Choi 2007)(TAM and AI, Yoko SASAKI, Satoshi KAGAMI)(Valin, Rouat, and Michaud 2004)(Yoshida, Kazuhiro Nakadai, and Hiroshi G. Okuno 2009). Additionally, this modern audio collection system is starting to be used more and more in videogames station (Chetty 2009), laptops, mobile phones, cars (Oh, Viswanathan, and Papamichalis 1992)… In almost all, using a non commercial microphone array with 3 or 4 unidirectional microphones, and over that, noise cancellation and a source location algorithms are applied by software.

This kind of microphones is very robust to noise, and combines the advantages of using unidirectional microphones (not very affected by background noise) and of using omnidirectional microphones (you can speak without using headphones or "earpiece": freedom of movement and headset quality without the headset).

Fig. 9 Multi-array microphone

Recently, the first general-purpose commercial microphone arrays are starting to make their way into the market. They are endowed with signal processing algorithms for noise cancelling and source location by hardware. These devices are still relatively expensive. However, we have acquired a commercial eight microphone array device[14] and tested it once built into the robot Maggie. See Fig. 9 for a graphical representation of results.

Again the test have been performed at different distances to the robot (1, 2 and 3 or more meters) without important background noise (less than 50 dB). The results are:

1m: **0.62** confidence value (95% success rate).
2m: **0.53** confidence value (83.78% success rate).
>3m: **0.37** confidence value (52.5% success rate)

Results at the same test scenarios but with important noise background (about 65 dB) are:

1m: **0.47** confidence value (81,08 % success rate).
2m: **0.41** confidence value (80 % success rate).
>3m: **0.31** confidence value (31,67 % success rate)

The fact that the recognition's accuracy decreases as we move away from the microphone can be appreciated. This also happens when we add significant ambient noise. In the next section, we will compare these results with the previous tests, taken with unidirectional and omnidirectional microphones.

SUMMARY OF ALL TEST SCENARIOS WITH DIFFERENTS MICROPHONES

To summarize the performed survey, we depict the three compared audio capture systems we used for the tests on a single graph (see Fig. 10). In this figure, the three audio capture systems' performance is compared in two environments: one quiet (less than 50 dB), and one noisy (approximately 65 dB). Sound is captured from a distance equal to or less than 2 meters from the user's mouth (in the case of the headset, it is actually extremely close).

With these results, the fact that using the headset provides the most precise values in speech recognition can be observed. This system is followed in rank by the microphone array system, and, finally, the omnidirectional microphone. In all three cases, we have used high-end professional microphones (highest quality based on currently available devices on the market as of 2010, with the exception of the microphone array system: a new model with noise cancellation hardware has been released recently and has still not been acquired by the research group; it is said to be able to greatly improve the results of its category compared to its predecessors[15]).

Depending on the type of interaction and environment, it is desirable to use a microphone array integrated into the robot, or a headset placed next to the user's mouth. If the environment is very noisy and/or high recognition accuracy is very necessary, the headset is to be used. If the environment is quieter, and focus is on having a more natural interaction and/or sound source location, the microphone array built into the robot is to be used.

[14] http://www.acousticmagic.com/voice-tracker-array-microphone-technology.html
[15] http://www.acousticmagic.com/voice-tracker-ii-array-microphone-product-details.html.

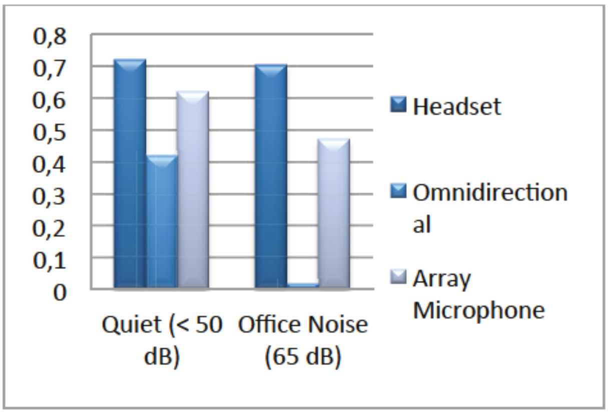

Fig. 10 Accuracy recognition comparative microphones (confidence values).

VII. HOW CLASSIFY ASR OUTPUTS AS ACCEPTED O REJECTED (A NEW APPROACH)

Usually programmer uses the confidence scores in order to decide whether to accept or reject asr outcomes. These classifications can reduce the incidence of misunderstanding but these require thresholds to be set which are themselves notoriously difficult to optimize. Modern recognizers can produce recognition hypotheses but it is not clear in practice how these can be used effectively.

As we said, most spoken dialog systems (SDS) are based in this local use of confidence and they need a confidence threshold. Some improvements allow them to adapt dynamically this threshold or change the confirmation strategy (implicit or explicit confirmation request). Instead other smarter SDS POMPD-based (Roy, Pineau, and Thrun 2000) no need use a threshold since they maintain a distribution across all states rather than a point-estimate of the most likely state, and SDS track all possible dialogue paths rather than just the most likely path, but the use of this systems for any practical system is, however, far from straightforward.

In this work we propose a new way to use the ASR results to accept or reject them. This new approach follow using the confidence score but in a more intelligent way; we have called it "second opinion". For this we have developed a confidence annotation component, which uses features form different knowledge sources in the system to compute a new confidence score more reliable.

"Second opinion" is based on using several recognition engines at the same time (at least two). In this work, we have used two recognition engines in parallel, the first of them is Loquendo ASR, outlined above, and the second recognition engine have been Google Voice ASR (it can be tested with Chrome HTML5, Android SDK and Youtube Automatic Subtitles), but you can use any other. We have tested with Google ASR because it is an online recognition engine and therefore it does no consume local CPU.

In a typical interaction, Loquendo ASR provides the audio log files that we can send to online Google ASR. Audio files only have got voice samples (additional samples are deleted for Loquendo Speak Detector). Google ASR processes these audio files and return to the local application the ASR results. These Loquendo and Google ASR outcomes can be processed for the new confidence annotation component to calculate a new confidence score.

To calculate this new average confidence score, firstly we have used the next function:

C1 = results confidence ASR1(it is provided for each recognition by the first ASR engine and is used to show the guarantee that the recognizer has performed correctly).
C2 = results confidence ASR2 (it is provided for each recognition by the second ASR engine and is used to show the guarantee that the recognizer has performed correctly).
SNR1 = Signal to Noise Ratio ASR1 (it is provided for each recognition by the fisrt ASR engine and is used to show, in each recognition, the noise level environment).

SNR2 = Signal to Noise Ratio ASR2 (it is provided for each recognition by the second ASR engine and is used to show, in each recognition, the noise level environment).

SR1 = pre-calculated success rate ASR1 (success rate obtained testing in real environments by the first ASR engine and is used to show percentage of times that the first recognizer is capable of matching the correct sentence).

SR2 = pre-calculated success rate ASR2 (success rate obtained testing in real environment by the second ASR engine and is used to show percentage of times that the second recognizer is capable of matching the correct sentence).

***AC** = average-confidence (final confidence taking account the relative weight of each recognizer).*

$$\text{(I) } AC = \left(\frac{SR1}{SR1+SR2}\right) * C1 + \left(\frac{SR2}{SR1+SR2}\right) * C2$$

With this function we give more weight/value to the recognizer with more a priori known success rate. We have calculated success rate testing with the engines and it is showed in “Table 1 ASR Framework comparison table”.

In the previous equation, we have not taken into account that the success rate for each recognition engine is strongly influenced by noisy, thus the success rate should not be a uniform function for all noise values. Probably a recognizer is more affected by noise that another, and the results provided by this recognizer more sensitive in noise ambient should be less reliable than the results provided for another more robust engine against ambient noise. Therefore is logical penalize the noise-sensitive engine in noise environments and reward it in silent environments. Therefore we need calculate a function that relation the SNR with the success rate for each recognizer, how we did in the Fig. 6. Another example about this function is in Fig. 11.

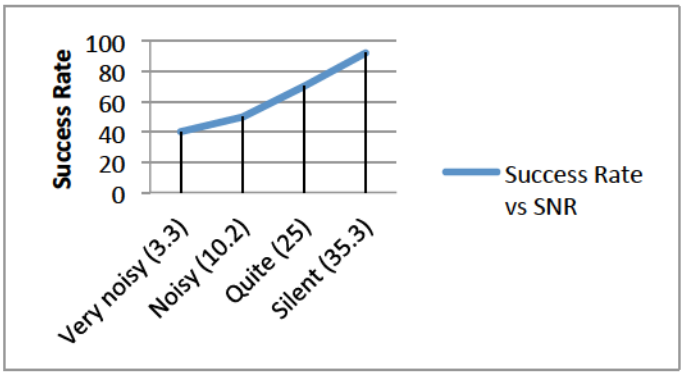

Fig. 11 Success Rate vs SNR: success rate probability function (PSR)

SNR score can be obtained in each recognition and hence we can build that function with a lot of recognitions, and after use this one for calculate a new average confidence as we formulating in function (II). Remark that SNR score it is influence by several factors as model and type of microphone, reverberation and, of course, noise environment.

$$\text{(II) } AC = \frac{PSR1(SNR1)}{PSR1(SNR1)+PSR2(SNR1)} * C1 + \frac{PSR2(SNR2)}{PSR1(SNR2)+PSR2(SNR2)} * C2$$

In Fig. 12 we have compared the average-confidence using equation (II) and the traditional confidence using only one recognizer, Loquendo, and the threshold was changed between 0 and 1, the two axes in the curve are false alarm and missing, which are defined as:

$$\text{(III) } False\ Alarm = \frac{FP}{FP+TN}$$

$$(IV)\ Missing \ = \frac{FN}{TP+FN}$$

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Tabla 2 Definition table

Basically in X axe we represented the accepted utterance that were wrong and in Y axe we represented the utterance was rejected were rights. To understand the figure, see Tabla 2 and equation (3) and (4).

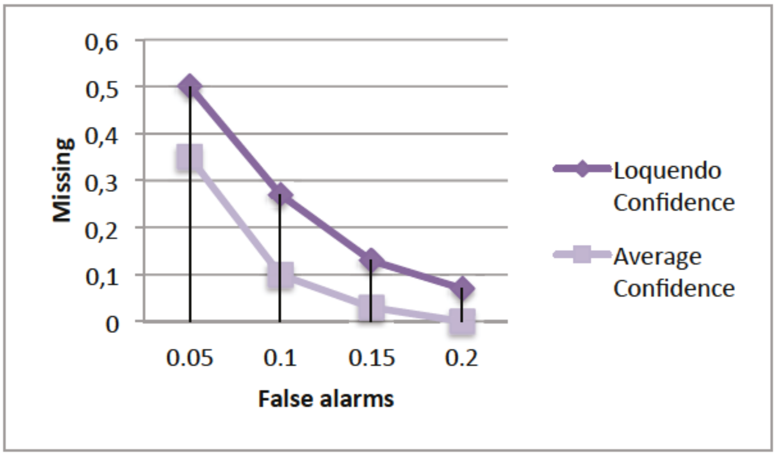

Fig. 12 Average Confidence vs typical confidence

When the value of false alarm is high, i.e. a lot of wrong utterance are accepted, involves that the missing value is low, that is, a few right utterance are rejected because most utterance (good and bad) are accepted. Instead when the false alarm is low the missing value is high because many utterance are rejected. A correct combination of confidence score calculation and threshold, try to maintain a right trade-off between false alarm and missing, decreasing missing value and false alarms. With this work we achieve this.

VIII. CONCLUSIONS

We have presented our work and the steps taken to give a social robot the capability of understanding natural spoken language as precisely as possible. We have focused on analyzing the best commercial recognition engines with their advantages and disadvantages, and to choose which most suits our needs and provides us with more potential. We have analyzed the accuracy of speech recognizer in many possible environments and situations. We have additionally focused on choosing the most suitable system to capture audio and improve human-robot interaction, providing complete freedom of movement and natural language.

In this work, we have showed the steps, requirements, frameworks, hardware and how integrate all in any robot, thus this work gives guidelines and advice in how to incorporate automatic speech recognition on a generic robotic platform, but the paper not only aims to provide a recent survey of existing ASR technologies and microphones. We have also focused in test some parameters in robot audition as user speaker volume, user intonation, user age and sex, separation between speakers and robot and real influence among noise and recognition accuracy. It is important remark that it is expensive and laborious to get user license of this commercial software and hardware. Once that it is achieved, to configure, install, test and to integrate these systems in a robotic platform requires a major effort through this work we try to alleviate.

We have verified that array-microphone systems still do not work so well in noisy environments as directional microphones do, even though the commercial array-microphone tested claim better results that directional microphones. The major problem is that this topic is yet under active research, and thus commercially hardware available system does not provide any better solution about feasibility of ASR to social robot that directional microphones do.

Finally we have presented a new method to classify ASR results as accepted or rejected based in a "second opinion" that decreases the number of false positives and also decreases the false negatives considering the pre-calculated success-rate for each recognition engine in each SNR interval. To apply this method it is necessary to work with, at least, two ASR engines in parallel and pre-calculate the success rate in noise intervals for each.

IX. ACKNOWLEDGMENTS

The authors gratefully acknowledge the funds provided by the Spanish Government through the project called `Peer to Peer Robot-Human Interaction" (R2H), of MEC (Ministry of Science and Education), and the project "A new approach to social robotics" (AROS), of MICINN (Ministry of Science and Innovation). The research leading to these results has received funding from the RoboCity2030-II-CM project (S2009/DPI-1559), funded by Programas de Actividades I+D en la Comunidad de Madrid and cofunded by Structural Funds of the EU.

X. REFERENCS

Barber, R., and Ma Salichs. 2002. A new human based architecture for intelligent autonomous robots. In *Intelligent autonomous vehicles 2001 (IAV 2001): a proceedings volume from the 4th IFAC Symposium, Sapporo, Japan, 5-7 September 2001*, 81. Pergamon. http://scholar.google.es/scholar?cluster=10839062160608396845&hl=es&as_sdt=2000#0.

Bennett, Christina, A.F. Llitjós, Stefanie Shriver, A.I. Rudnicky, and Alan W Black. 2002. Building VoiceXML-based applications. In *Seventh International Conference on Spoken Language Processing*, 2-5. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.228&amp;rep=rep1&amp;type=pdf.

Bischoff, R., and V. Graefe. 2004. HERMES - a versatile personal robotic assistant. *Proceedings of the IEEE* 92, no. 11: 1759-1779. doi:10.1109/JPROC.2004.835381. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1347457.

Breazeal, Cynthia. 2003. Emotive qualities in lip-synchronized robot speech. *Advanced Robotics* 17, no. 2 (May): 97-113. doi:10.1163/156855303321165079. http://www.ingentaconnect.com/content/vsp/arb/2003/00000017/00000002/art00003.

Chan, S.W.K. 1995. Inferences in natural language understanding. In *Proceedings of 1995 IEEE International Conference on Fuzzy Systems. The International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium*, 935-940. IEEE. doi:10.1109/FUZZY.1995.409794. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=409794.

Chetty, Vasu. 2009. Microsoft's Project Natal for Xbox 360 (July). http://www.suite101.com/content/microsofts-project-natal-for-xbox-360-a129412.

Dalmasso, E., F. Castaldo, P. Laface, D. Colibro, and C. Vair. Loquendo - Speaker recognition evaluation system. In *Acoustics, Speech and Signal Processing, ICASSP 2009. IEEE International Conference on*, 4213-4216. Taipei. doi:10.1109/ICASSP.2009.4960558. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4960558.

Favre, Benoit, Bernd Bohnet, and Dilek Hakkani-Tur. 2010a. Evaluation of semantic role labeling and dependency parsing of automatic speech recognition output. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing* 1: 5342-5345. doi:10.1109/ICASSP.2010.5494946. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5494946.

---. 2010b. Evaluation of semantic role labeling and dependency parsing of automatic speech recognition output. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5342-

5345. Dallas (USA): IEEE. doi:10.1109/ICASSP.2010.5494946. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5494946.

Fry, J., H. Asoh, and T. Matsui. 1998. Natural dialogue with the Jijo-2 office robot. In *Intelligent Robots and Systems, 1998. Proceedings., 1998 IEEE/RSJ International Conference on*, 1278-1283. Victoria (Canada). doi:10.1109/IROS.1998.727475. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=727475.

Gorostiza, Javi, Ramon Barber, Alaa Khamis, Maria Malfaz, Rakel Pacheco, Rafael Rivas, Ana Corrales, Elena Delgado, and Miguel Salichs. 2006. Multimodal Human-Robot Interaction Framework for a Personal Robot. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 39-44. Hatfield (UK): IEEE, September. doi:10.1109/ROMAN.2006.314392. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4107783.

Hegel, Frank, Claudia Muhl, Britta Wrede, Martina Hielscher-Fastabend, and Gerhard Sagerer. 2009. *Understanding Social Robots. 2009 Second International Conferences on Advances in Computer-Human Interactions*. IEEE, February. doi:10.1109/ACHI.2009.51. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4782510.

Huang, J, Noboru Ohnishi, and Noboru Sugie. 1997. Building ears for robots: Sound localization and separation. *Artificial Life and Robotics* 1, no. 4 (December): 157-163. doi:10.1007/BF02471133. http://www.springerlink.com/content/upw68k6138152679/.

Huang, XD, KF Lee, H.W. Hon, and M.Y. Hwang. 1991. Improved acoustic modeling with the SPHINX speech recognition system. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, 345-348. Toronto (Canada): IEEE. doi:10.1109/ICASSP.1991.150347. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=150347.

Ishi, Carlos, Shigeki Matsuda, Takayuki Kanda, Takatoshi Jitsuhiro, Hiroshi Ishiguro, Satoshi Nakamura, and Norihiro Hagita. 2006. Robust Speech Recognition System for Communication Robots in Real Environments. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, 340-345. Genoa (Italy): IEEE, December. doi:10.1109/ICHR.2006.321294. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4115624.

Jansen, B, and T Belpaeme. 2006. A computational model of intention reading in imitation. *Robotics and Autonomous Systems* 54, no. 5 (May): 394-402. doi:10.1016/j.robot.2006.01.006. http://linkinghub.elsevier.com/retrieve/pii/S0921889006000194.

Junlan, Feng. 2010. A general framework for building natural language understanding modules in voice search. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5362-5365. Dallas: IEEE, March. doi:10.1109/ICASSP.2010.5494951. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5494951.

Kibria, Shafkat, and T. Hellström. 2007. Voice user interface in robotics - common issues and problems. *aass.oru.se*. http://www.aass.oru.se/Research/Learning/publications/2007/Kibria_Hellstrom_2007-CS07-Voice_User_Interface_in_Robotics_Common_Issues_and_Problems.pdf.

Kim, H, and Js Choi. 2007. Human-robot interaction in real environments by audio-visual integration. *INTERNATIONAL JOURNAL OF CONTROL* 5, no. 1: 61-69. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.8455&amp;rep=rep1&amp;type=pdf.

Kim, K, Minwoo Jeong, and GG Lee. 2007. Improving Speech Recognition Using Semantic and Reference Features in a Multimodal Dialog System. In *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 416-420. Jeju Island (Korea): IEEE. doi:10.1109/ROMAN.2007.4415120. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4415120.

Lecouteux, Benjamin, Pascal Nocera, and Georges Linares. 2010. *Semantic cache model driven speech recognition*. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. doi:10.1109/ICASSP.2010.5495642. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5495642.

Li, J, and Kaizhu Wang. 1993. Natural language understanding based on background knowledge. In *Proceedings of TENCON '93. IEEE Region 10 International Conference on Computers, Communications and Automation*, 460-462. Beijing: IEEE. doi:10.1109/TENCON.1993.320026. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=320026.

Lin, Feng, and Fuliang Weng. 2010. Computing confidence score of any input phrases for a spoken dialog system. In *2010 IEEE Spoken Language Technology Workshop*, 295-300. Berkeley, California (USA): IEEE, December. doi:10.1109/SLT.2010.5700867. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5700867.

Llisterri, Joaquim, Carme Carbó, Mj Machuca, C. De la Mota, M. Riera, and A. R'\ios. 2003. El papel de la fonética en el desarrollo de las tecnologías del habla. In *Memorias de las VII Jornadas de Linguística.*

Cadiz (Spain): Servicio de Publicaciones de la Universidad de Cádiz. http://liceu.uab.es/~joaquim/speech_technology/UNAM_03/UNAM03_Guion_Bib.pdf.

Lopes, Luis Seabra, and Tony Belpaeme. 2008. Beyond the individual: new insights on language, cognition and robots. *Connection Science* 20, no. 4 (December): 231-237. doi:10.1080/09540090802518661. http://www.informaworld.com/openurl?genre=article&doi=10.1080/09540090802518661&magic=crossref||D404A21C5BB053405B1A640AFFD44AE3.

Mitsunaga, N., T. Miyashita, H. Ishiguro, K. Kogure, and N. Hagita. 2006. Robovie-IV: A Communication Robot Interacting with People Daily in an Office. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 5066-5072. Beijing: IEEE. doi:10.1109/IROS.2006.282594. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4059225.

Nakadai, Kazuhiro, Tino Lourens, Hiroshi G. Okuno, and Hiroaki Kitano. 2000. Active Audition for Humanoid. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, 832–839. AAAI Press. http://portal.acm.org/citation.cfm?id=647288.723417.

Nakadai, Kazuhiro, Hiroshi G Okuno, Hirofumi Nakajima, Yuji Hasegawa, and Hiroshi Tsujino. 2008. An Open Source Software System For Robot Audition HARK and Its Evaluation.

Niklfeld, Georg, and Robert Finan. 2001. Architecture for adaptive multimodal dialog systems based on VoiceXML. In *Proceedings of EuroSpeech*, 1-4. Scandinavia: Association for Computational Linguistics Morristown. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.7658&amp;rep=rep1&amp;type=pdf.

Nyberg, Eric, Teruko Mitamura, Paul Placeway, Michael Duggan, and San Francisco. 2002. DialogXML: Extending VoiceXML for Dynamic Dialog Management. In *Proceedings of the second international conference on Human Language Technology Research*, 298-302. San Diego (California): Morgan Kaufmann Publishers Inc.

Oh, S., V. Viswanathan, and P. Papamichalis. 1992. Hands-free voice communication in an automobile with a microphone array. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 281-284. San Francisco: IEEE. doi:10.1109/ICASSP.1992.225916. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=225916.

Okuno, Hiroshi G. 2007. Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. In *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 111-116. Kyoto (Japan): IEEE. doi:10.1109/ASRU.2007.4430093. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4430093.

Paolo Baggia, Silvia Mosso. 2005. Loquendo Speech Technologies and multimodality.

R. Rivas, A. Corrales, R. Barber, M. A. Salichs. 2007. Robot Skill Abstraction for AD Architecture. *6th IFAC Symposium on Intelligent Autonomous Vehicles*. http://roboticslab.uc3m.es/publications/iav07_AD.pdf.

Roy, Nicholas, Joelle Pineau, and Sebastian Thrun. 1998. Spoken Dialog Management for Robots. *Management*.

---. 2000. *Spoken dialogue management using probabilistic reasoning. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL '00*. Morristown, NJ, USA: Association for Computational Linguistics, October. doi:10.3115/1075218.1075231. http://portal.acm.org/citation.cfm?id=1075218.1075231.

Sakagami, Y., R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, and K. Fujimura. 2002. The intelligent ASIMO: system overview and integration. In *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, 2478-2483. Laussane: IEEE. doi:10.1109/IRDS.2002.1041641. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1041641.

Shuyin, Ioannis Toptsis, Ioannis Toptsis, S Li, Britta Wrede, and Gernot A. Fink. 2004. A Multi-modal Dialog System for a Mobile Robot. In *Int. Conf. on Spoken Language Processing*, 273-276. Jeju Island (Korea): IEEE. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.59.9237.

Sofge, Donald, J Gregory Trafton, Nicholas Cassimatis, Dennis Perzanowski, Magdalena Bugajska, William Adams, and Alan Schultz. Human-Robot Collaboration and Cognition with an Autonomous Mobile Robot. *Artificial Intelligence*.

Takahashi, T., K. Nakadai, K. Komatani, T. Ogata, and H.G. Okuno. 2010. Improvement in listening capability for humanoid robot HRP-2. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 470-475. Anchorage: IEEE. doi:10.1109/ROBOT.2010.5509830. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5509830.

TAM, Yuki, and Hiroshi Mizoguchi AI, Yoko SASAKI, Satoshi KAGAMI. Three Ring Microphone Array for 3D Sound Localization and Separation for Mobile Robot Audition.

Tamai, Y., Y. Sasaki, S. Kagami, and H. Mizoguchi. 2005. Three Ring Microphone Array for 3D Sound Localization and Separation for Mobile Robot Audition. In *IEEE/RSJ International Conference on*

Intelligent Robots and Systems, 903-908. Edmonton (Canada): IEEE. doi:10.1109/IROS.2005.1545095. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1545095.

Tanaka, Nobuaki, Tetsuji Ogawa, Kenzo Akagiri, and Tetsunori Kobayashi. 2010. DEVELOPMENT OF ZONAL BEAMFORMER AND ITS APPLICATION TO ROBOT AUDITION. In *Signal Processing*, 1:1529-1533. http://www.eurasip.org/Proceedings/Eusipco/Eusipco2010/Contents/papers/1569292345.pdf.

Valin, J.-M., J. Rouat, and F. Michaud. 2004. Enhanced robot audition based on microphone array source separation with post-filter. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2123-2128. Sendai (Japan): IEEE. doi:10.1109/IROS.2004.1389723. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1389723.

---. *Enhanced robot audition based on microphone array source separation with post-filter. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (IEEE Cat. No.04CH37566)*. IEEE. doi:10.1109/IROS.2004.1389723. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1389723.

Valverde-Albacete, F.J., and J.M. Pardo. 1996. A multi-level lexical-semantics based language model design for guided integrated continuous speech recognition. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, 224-227. Philadelphia (USA): IEEE. doi:10.1109/ICSLP.1996.607082. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=607082.

Walker, D.E. 1976. Speech Understanding Through Syntactic and Semantic Analysis. *IEEE Transactions on Computers* C-25, no. 4 (April): 432-439. doi:10.1109/TC.1976.1674625. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1674625.

Wallis, Peter. 2010. A robot in the kitchen (July 15): 25-30. http://portal.acm.org/citation.cfm?id=1870559.1870564.

Yoshida, Takami, Kazuhiro Nakadai, and Hiroshi G. Okuno. 2009. Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, 604-609. Paris (France): IEEE, December. doi:10.1109/ICHR.2009.5379586. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5379586.

Yu, Xingang, Faguo Zhou, Fan Zhang, and Bingru Yang. 2009. Intelligent Decision Support System Based on Natural Language Understanding. In *2009 International Conference on Management and Service Science*, 1-4. Beijing (China): IEEE, September. doi:10.1109/ICMSS.2009.5302806. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5302806.