# Audition: the most important sense for humanoid robots?

*Rodolphe Gelin* [1], *Gabriele Barbieri* [1]

[1] A-Lab, Aldebaran Robotics, Paris, France

rgelin@aldebaran-robotics.com, ggabrieli@aldebaran-robotics.com

## Abstract

In this paper, we introduce NAO, the humanoid robot developed by Aldebaran Robotics. For this robot, dedicated to be a companion, the interaction with humans is crucial and the audition appears to be a very important sense while the robotic community has been focusing on vision for a long time. Some of the developments on audio are presented.

**Index Terms**: humanoid, robotics, audition

## 1. Introduction

Created in 2005, Aldebaran Robotics designs, develops and manufactures the humanoid robot NAO. Today more than 3000 NAOs have been sold all over the world for research and education purposes. If NAO appears to be a very efficient and appreciated development platform for these academic markets, the final objective of the company is to make humanoid robot a real companion for domestic applications. As soon as the robot is integrated as an autonomous assistant at home, its interface with its user becomes essential. Aldebaran strongly believes that the audition of the robot is crucial: it is a very intuitive way for communication, it can be omnidirectional and it can detect human or non-human events much before the vision can. That is the reason why Aldebaran has been and will be involved in many research projects about audition.

## 2. Presentation of the robot

NAO is a 58cm high humanoid robot equipped with 25 joints. It is able to walk, to make expressive gestures and to grasp simple objects. Its appearance has been designed to create the desire of interaction as soon as you see it. Beyond the sensors dedicated to motion (joint sensors, inertial unit, pressure sensors …), it is equipped with many devices that can carry this interaction with human: 2 cameras, 4 microphones, 2 loudspeakers in the head, colored LEDS in the eyes and tactile sensors on the head and on the hands. Three of our 5 senses (vision, audition, touch) can be used to interact with NAO.

Figure 1: *NAO interacting with its user.*

For the audio interaction, the embedded software provides text-to-speech and automatic speech recognition features that make NAO able to speak and listen in 8 different languages. The use of interaural time difference between the four microphones allows sound localization that is essential when the user willing to interact with NAO is not in front of it and cannot be detected by the cameras. Based on these features, NAO can propose, for instance, "audio menus" in which it can explain what it can do and listen to the choice of it user.

But the use of this robot is a domestic environment by non-expert users in real situation brings specific difficulties: the speaker is often far from the microphones of the robot, the environment is reverberating and noisy and the robot itself is a source of noises: mechanical noises when it moves, vibrations due to the embedded fan and the own voice of the robot is a very perturbing noise for the its microphones. To make the audio sense really usable a lot of work has still to be done and when the audio signal is available a lot of processing is possible to make the interaction ever better.

## 3. Current development

Within several internal and collaborative projects, Aldebaran has been working on the improvement of the audio acquisition. In partnership Telecom Paristech, Aledebaran is working on a dedicated DSP board that could provide a first level of audio processing (noise filtering, source sharing, echo cancellation) to bring better quality acoustic signal to the CPU while relieving its computing load. This principle of smart sensor, carrying a part of the dedicated processing, will be probably generalized on the robots. Within the ECHORD experiment BABIR, the French company Vocally worked with Aldebaran on smart noise filtering: during a first learning phase, the robot is able to learn its own mechanical noises to cancel them from the audio signal kept by the microphones. The same principle had been tried for echo cancellation but as soon as the voice of the robot is too loud, it saturates the microphones and the cancellation becomes impossible. A new physical implementation of the microphones and the loudspeakers will be necessary. A new positioning of the microphones on the head has been tested and the first results are very promising to lower the impact of the internal noise of the robot and to improve the directivity of the audio detection by implementing beam forming algorithms.

Based on the improved audio signal coming from this first level of processing, higher level features can be implemented. Within the Défi Carotte, organized by ANR and DGA, Aldebaran has worked with the French company Voxler on the sound recognition feature. The robot is able to recognize domestic sounds that can be used to trigger relevant behaviors (child's crying, breaking glass, phone ring, doorbell…). If the results are rather good when the audio source is not too far from the robot, as soon as the distance grows, the detection is worse and the results are difficult to exploit. Within the Romeo project, CNRS LIMSI has pretty good results in recognizing emotions in the voice of the speaker. Four basic emotions can be detected (neutral, joy, sadness, angry).

## 4. A dialog with NAO

NAO is able to interact verbally with humans, by listening and talking with them. Indeed, the NAO's dialog engine is intended to give to the robot the ability to interact and have a discussion with humans, by elaborating audio inputs and giving appropriate answers to them.

During a conversation, the audio coming from the person talking to NAO is processed from the text-to-speech module and converted to a sequence of characters. The dialog engine treats this input sequence by linking it to the appropriate output sequence. The output will finally be converted to animated speech.

Dialog context dramatically improve audio recognition. In fact NAO's speech-to-text module exploits an embedded language model based on the contextual Dialog. This embedded model is proved to be more robust than a general purpose language model.

During the conversation, NAO exploits all the benefits to being a humanoid robot. The audio input is enriched by input signals received by all the others input sense (vision and touch), to have a dialog based on actual situation, qualitative better than "abstract" dialog. Moreover, NAO speech outputs are enriched with body talking, exactly as a human will do, creating a stronger interaction.

## 5. Conclusions

From the lowest level of signal processing to the abstract dialog, from the smart sensors to the fusion with the vision, the exploitation of the audio signal will be for long time a research and development area for Aldebaran. Its teams in this domain are still growing and the collaboration with the best laboratories will be necessary to keep the leading position that Aldebaran is targeting for the mass market of humanoid robotics.

## 6. Acknowledgements

The developments presented in this paper have been supported by the 7th Frame Program of the European Community, by the Agence Nationale de la Recherche, the French Minister of Industry, the Region Ile de France and the City of Paris.