NAO

KEY FEATURE
SOUND SOURCE LOCALIZATION

Abstract

One of the main purposes of having a humanoid robot is to have it interact with people. This is undoubtedly a tough task that implies a fair amount of features. Being able to understand what is being said and to answer accordingly is certainly critical but in many situations, these tasks will require that the robot is first in the appropriate position to make the most out of its sensors and to let the considered person know that the robot is actually listening/talking to him by orienting the head in the relevant direction. The "Sound Localization" feature addresses this issue by identifying the direction of any "loud enough" sound heard by NAO.

Related work

Sound source localization has long been investigated and a large number of approaches have been proposed. These methods are based on the same basic principles but perform differently and require varying CPU loads. To produce robust and useful outputs while meeting the CPU and memory requirements of our robot, the NAO's sound source localization feature is based on an approach known as "Time Difference of Arrival".

Principles

The sound wave emitted by a source close to NAO is received at slightly different times on each of its four microphones. For example, if someone talks to the robot on his left side, the corresponding signal will first hit the left microphones, few milli-seconds later the front and the rear ones and finally the signal will be sensed on the right microphone (**FIGURE 1**).

These differences, known as ITD standing for "interaural time differences", can then be mathematically related to the current location of the emitting source. By solving this equation every time a noise is heard the robot is eventually able to retrieve the direction of the emitting source (azimutal and elevation angles) from ITDs measured on the 4 microphones.

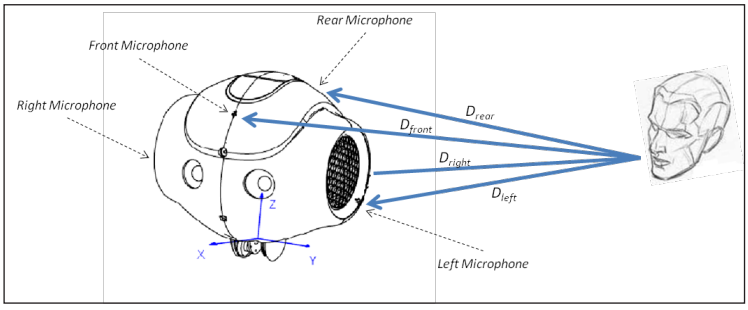

FIGURE 1
Schematic view of the dependency between the position of the sound source (a human in this example) and the different distances that the sound wave need to travel to reach the four NAO's micro-phones. These different distances induce times differences of arrival that are measured and used to compute the current position of the source.

Performances

The angles provided by the NAO's sound source localization engine match the real position of the source with an average accuracy of 20 degrees, which is satisfactory in many practical situations. Note that the maximum theoretical accuracy depends on the microphones' spatial configuration and on the sample rate of the measured signal, and is about 10 degrees on NAO.

The distance separating NAO and a sound source successfully located can reach several meters depending on the situation (reverberation, background noise, etc...).

Once launched, this feature uses 10% of the CPU constantly and up to 20% for few milliseconds when the location of a sound is being computed.

Limitations

The performance of NAO's sound source localization is limited by how clearly the sound source can be heard with respect to background noise. Noisy environments naturally tend to decrease the reliability of the module outputs.

It will also detect and locate any "loud sounds" without being able by itself to filter out sound source that are not humans.

Finally, only one sound source can be located at a time. The module can behave in a less reliable manner if NAO faces several loud noises at the same time. He will likely only output the direction of the loudest source.

How does it work?

This feature is available as a NaoQi module named "ALAudioSourceLocalization" which provides a C++ and Python API (application programming interface) that allows precise interactions from a python script or a NaoQi module.

Two boxes in Choregraphe are also available that allow an easy use of the feature inside a behavior:

- The box "Sound Loc." provides the output (angles and level of confidence) of the sound localization module without taking any further actions.
- The box "Sound Tracker" uses these outputs to make NAO's head turn in the appropriate direction.

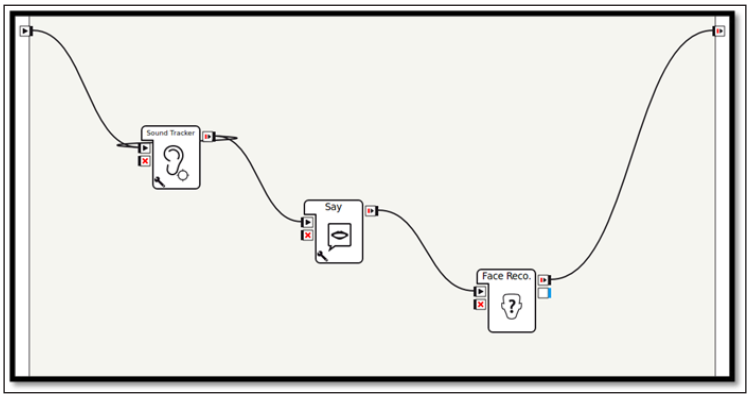

FIGURE 2
An example of a behavior using the «Sound Tracker» box to orientate NAO's head so that the «Face Recognition» box can actually perform its recognition.

How are people using it?

Here are some possible applications (from the simplest to the more ambitious ones) that can be built from NAO's ability to locate sound sources.

- Using the "Sound Source Localization" to have a person enter the camera field of view (as shown in the above example). This allows subsequent vision based features to work on relevant images (images showing a person for example). This is consequently of interest for these specific tasks:

- Human Detection, Tracking and Recognition
- Noisy Objects Detection, Tracking and Recognition

- "Sound Source Localization" can be used to strengthen the Signal/Noise ratio in a specific direction - this is known as Audio Source Separation – and can critically enhance subsequent audio based algorithms such as:

- Speech Recognition in a specific direction
- Speaker Recognition in a specific direction

- Theses possible applications can also be mixed together making NAO's sound source localization the basic block for sophisticated applications such as:

- Remote Monitoring / Security applications (NAO's could track noises in an empty flat, take pictures and record sounds in relevant directions, etc...)
- Entertainment applications (by knowing who speaks and understanding what is being said, NAO could easily take part in a great variety of games with humans.)