Binaural Hearing for Robots

Introduction to Robot Hearing

Radu Horaud

Binaural Hearing for Robots

1. **Introduction to Robot Hearing**
2. Methodological Foundations
3. Sound-Source Localization
4. Machine Learning and Binaural Hearing
5. Fusion of Audio and Vision

Radu Horaud

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. Human-robot interaction
3. Auditory scene analysis
4. Audio signal processing in brief
5. Audio processing in the ear
6. Audio processing in the midbrain
7. Audio processing in the brain

1. Introduction to Robot Hearing

1. **Why do robots need to hear?**
2. Human-robot interaction
3. Auditory scene analysis
4. Audio signal processing in brief
5. Audio processing in the ear
6. Audio processing in the midbrain
7. Audio processing in the brain

Robots in Industry

Researchers and engineers have developed sophisticated robots that can work efficiently in well-structured and predictable environments.

Examples: manufacturing industries (cars, ships, planes, electric appliances, etc.),

The emphasis has been on studying **physical interactions** between robots and objects/environment, for example:

- perform complex articulated movements,
- grasp/ungrasp and assemble/disassemble objects,
- move safely among obstacles,
- perform specific tasks: painting, welding, etc.

Robots and People

Robots are expected to gradually move from factory floors to populated spaces, therefore there is a paradigm shift from **robot-object interactions** to **human-robot interactions** (HRI).

HRI has the ambition to enable unconstrained communication between robots and people.

Robot hearing is a fundamental ability because it allows interaction via highly rich content, e.g., speech.

The challenge: Extract content from complex acoustic signals.

Acoustic Signals Are Very Rich

Robot hearing should not be limited to speech recognition.

There is a large number of acoustic events, other than speech, that are of interest to a robot:

- laughing, crying, whistling, singing, hand clapping, etc.
- animal sounds, musical instruments, environmental sounds,
- falling objects/people, warnings, electronic appliances, hazards, etc.

Session Summary

- Robots in industry
- Robots and people
- What can be done with hearing?

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. **Human-robot interaction**
3. Auditory scene analysis
4. Audio signal processing in brief
5. Audio processing in the ear
6. Audio processing in the midbrain
7. Audio processing in the brain

Constrained (Tethered) HRI

Inspired from human-computer interaction (HCI).

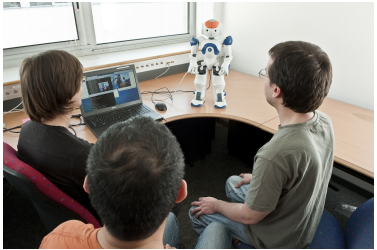
Advantage: clean acoustic signals available for automatic speech recognition (95% recognition rate).

Disadvantages:

- restricted to speech communication with a single user that must wear a microphone,
- sound localization and recognition of other types of acoustic events not possible.
- audio perception cannot be combined with other sensorial modalities (vision).

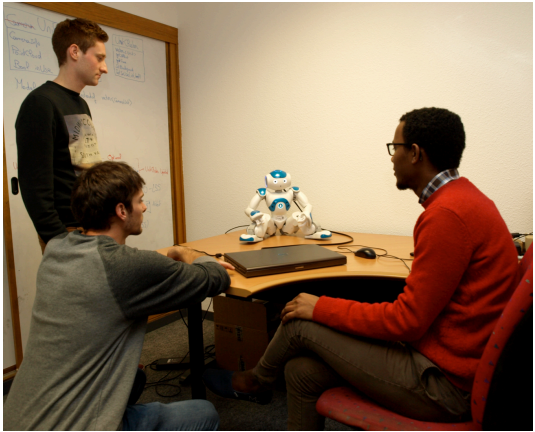
Unconstrained (Untethered) HRI

- The microphones are embedded into the robot head.
- The relevant auditory signals are at some distance from the robot.
- These signals are mixed with signals from other sound sources, affected by reverberations, noise, etc.
- Auditory processing is more difficult in this case.
- This setup allows auditory scene analysis.
- It is possible to combine audio and vision.



Automatic Speech Recognition

People-Robot Interaction



Session Summary

- Human-computer interaction
- Human-robot interaction
- Speech recognition
- Other hearing tasks

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. Human-robot interaction
3. **Auditory scene analysis**
4. Audio signal processing in brief
5. Audio processing in the ear
6. Audio processing in the midbrain
7. Audio processing in the brain

Auditory Scene Analysis

With two or more microphones it is possible to address several problems:

- detect the presence/absence of the sound sources,
- extract the signal emitted by each emitting source,
- localize the sounds and track them over time,
- identify each source (voice, musical instrument, etc.)
- enhance the signals for subsequent ASR.

Robot Hearing Challenges

A robot hearing system must solve several problems:

Analysis:

- Who are the speakers in front of the robot?
- Where are the speakers?
- When do they speak?
- What are they saying?

Interaction:

- Is the robot able to pop into the conversation?
- Is the robot able to synthesize appropriate behavior?

Constrained Audio Interaction

$s(t)$: the signal emitted by a speaker is a function of time t,

$m(t)$: the signal recorded with a microphone,

$n(t)$: additive noise.

We have:

$$m(t) = s(t) + n(t)$$

The emitted and recorded signals differ by additive noise (for example, microphone noise).

Extracting a **clean speech signal**, for subsequent automatic speech recognition (ASR), is straightforward in this case.

Unconstrained Audio Interaction

The relationship between the emitted and received signals is more complex

$$m(t) = h_1(t) \star s_1(t) + h_2(t) \star s_2(t) + \ldots + h_K(t) \star s_K(t) + n(t)$$

- Indeed, the microphone records not a single signal but several signals: $s_1(t), s_2(t), \ldots, s_K(t)$.
- When waves travel from their positions to the microphone, they are **modified**.
- These modifications can be modeled by **convolution** (the symbol $\star$) between an **unknown transfer function**, $h_k(t)$, and the emitted signal, $s_k(t)$.

Extracting **clean signals** is more difficult!

Session Summary

- Auditory scene analysis
- Robot hearing challenges
- Constrained interaction
- Unconstrained interaction

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. Human-robot interaction
3. Auditory scene analysis
4. **Audio signal processing in brief**
5. Audio processing in the ear
6. Audio processing in the midbrain
7. Audio processing in the brain

Using Several Microphones

For each microphone $j = 1 \ldots J$ we have a different equation:

$$
\begin{aligned}
m_1(t) &= h_{11}(t) \star s_1(t) + h_{21}(t) \star s_2(t) + \ldots + h_{K1}(t) \star s_K(t) + n_1(t) \\
&\vdots \\
m_j(t) &= h_{1j}(t) \star s_1(t) + h_{2j}(t) \star s_2(t) + \ldots + h_{Kj}(t) \star s_K(t) + n_j(t) \\
&\vdots \\
m_J(t) &= h_{1J}(t) \star s_1(t) + h_{2J}(t) \star s_2(t) + \ldots + h_{KJ}(t) \star s_K(t) + n_J(t)
\end{aligned}
$$

Binaural Hearing

The microphones are plugged into the left and right “ears" of a dummy head.

In the absence of reverberations, there is a **head-related transfer function** (HRTF) for each ear:

$$m_L(t) = h_L(t) \star s_1(t) + h_L(t) \star s_2(t) + \ldots + h_L(t) \star s_K(t) + n_L(t)$$
$$m_R(t) = h_R(t) \star s_1(t) + h_R(t) \star s_2(t) + \ldots + h_R(t) \star s_K(t) + n_L(t)$$

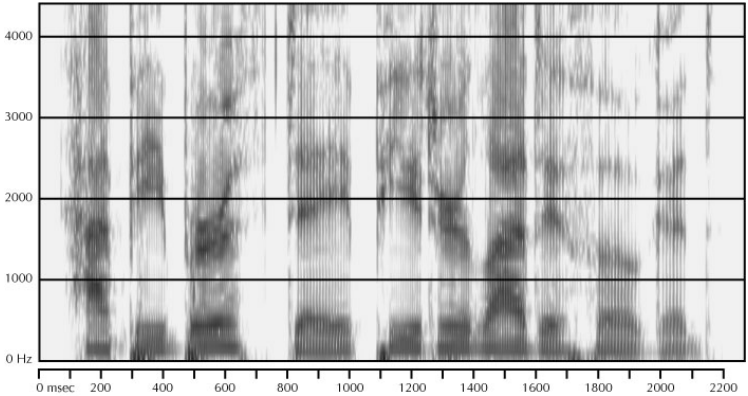
This general setting is difficult to solve in practice.

Spectral Representation

- In signal processing, it is common to perform spectral analysis, namely to apply the Fourier Transform (FT) to the microphone signals.
- In practice, one should use the discrete Fourier transform (DFT).
- The DFT applied over a short period of time is called the short-time Fourier transform (STFT).
- By applying the STFT to a signal at discrete time steps, one obtains a **spectrogram** (see next slide).
- Each spectrogram point indicates the amount of oscillation contained in the signal at time t and at frequency f

Spectrogram

A Spectrogram is an example of a time-frequency representation of a signal.



Binaural Cues for Sound Localization

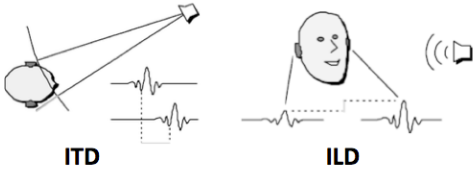
Suppose that there is a single emitting sound source.

The signals received by the two ears are **different**. Their difference is mainly characterized by two cues:

- They reach each ear at different times: interaural time difference, or ITD.
- They have different intensities (or levels): interaural intensity (or level) difference, or IID (ILD).

ITD and ILD

- ITD: interaural time difference (also TDOA or time difference of arrival)
- ILD: interaural level difference



- From these two features it is possible to localize the sound source.

Session Summary

- Audio signal processing
- Emitted and perceived signals
- Spectrograms
- Binaural localization cues

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. Human-robot interaction
3. Auditory scene analysis
4. Audio signal processing in brief
5. **Audio processing in the ear**
6. Audio processing in the midbrain
7. Audio processing in the brain

The Auditory Pathways: the Ears and the Brain

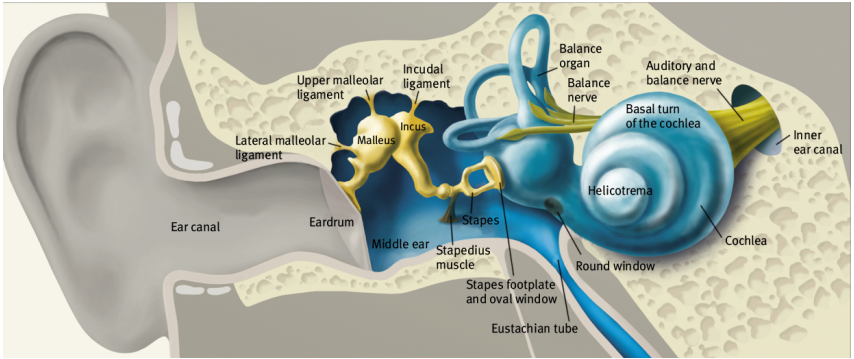
All the living species, from fish to primates (monkeys, humans) have binaural hearing.

There is considerable variability in the sub-cortical and cortical organization across species.

Roughly speaking, the auditory pathway can be divided into two parts:

- transformation of sound waves (air pressure) into spike trains (neural activity), and
- representation and extraction of auditory information (localization, speech, etc.) in different brain areas.

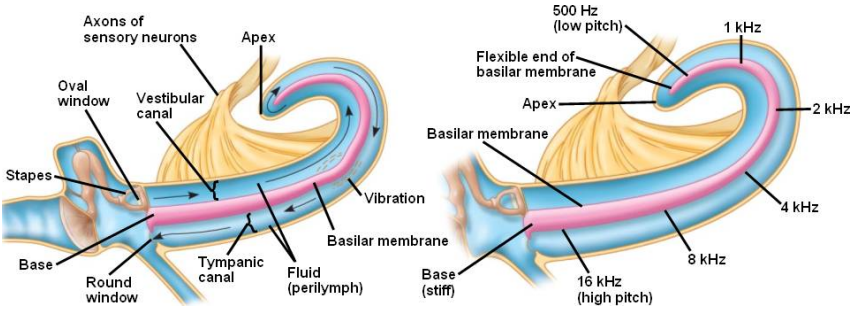
The Ear


Balance organ
Auditory and balance nerve
Upper malleolar ligament
Incudal ligament
Balance nerve
Basal turn of the cochlea
Incus
Lateral malleolar ligament
Malleus
Inner ear canal
Helicotrema
Ear canal
Eardrum
Stapes
Cochlea
Middle ear
Stapedius muscle
Round window
Stapes footplate and oval window
Eustachian tube

From Air Pressure to Neural Activity

- *Outer ear* : composed of the *pinna* and *ear canal*, transmits air to the *eardrum*.
- *Middle ear* : it is air filled, contains a chain of little bones, or *ossicles* that collect sound pressure on the *eardrum* an concentrates it onto the *cochlea*;
- *Cochlea* : Acts as a mechanical bank of bandwidth filters that transform the air/liquid vibrations into electric currents and then into spike trains.

The Cochlea


Axons of sensory neurons
Apex
Oval window
Vestibular canal
Stapes
Vibration
Basilar membrane
Base
Round window
Tympanic canal
Fluid (perilymph)
500 Hz (low pitch)
1 kHz
Flexible end of basilar membrane
Apex
2 kHz
Basilar membrane
4 kHz
8 kHz
Base (stiff)
16 kHz (high pitch)

The Main Components of the Cochlea

- A coiled tube enclosed in a hard bony shell,
- It is filled with a liquid (salted water),
- The *basilar membrane* splits the spaces inside the cochlea into two canals, vestibular and tympanic,
- The basilar membrane is stiff at its base and flexible at its apex,
- all along the basilar membrane there is a structure called *organ of Corti* that contains the *hair cells*,
- Hair cells are the hearing receptor cells, or the equivalent of photo-receptor cells on the retina.

The Cochlea: From Air Pressure to Electrical Signals

The cochlea is equipped with two types of mechanical resistance:

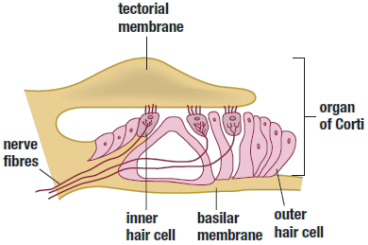
- the stiffness of the basilar membrane, and
- the inertia of the liquid

Both are graded along the cochlea: the stiffness gradient decreases while the inertial gradient increases.

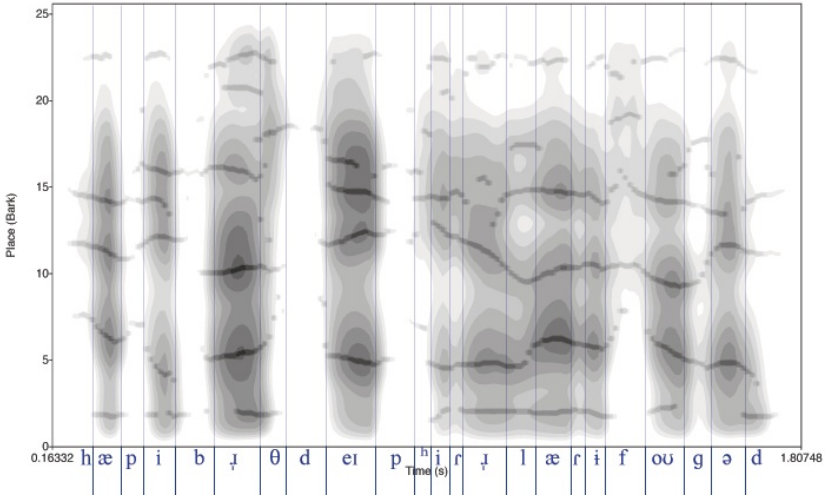
Altogether the cochlea operates as some kind of mechanical frequency analyzer, modeled as a *gamma tone filter bank*

The Organ of Corti

- There are 15000 hair cells all along the basilar membrane, inner and outer hair cells.
- The outer cells play the role of amplification.
- The inner hair cells transform mechanical vibrations into electric signals that are transmitted to the *auditory nerve*.
- Mathematically, the information passed to the brain can be represented as a **cochleargram** (see next slide) which is somehow equivalent to a spectrogram (see slide **??**)



Cochleargram



Session Summary

- The ear is a very complex organ
- Air pressure is transformed into electric spikes
- Detailed ear anatomy
- Brief description of ear's functions

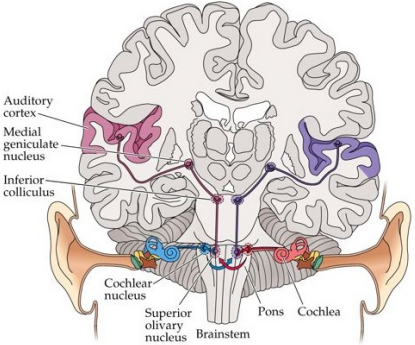
1. Introduction to Robot Hearing

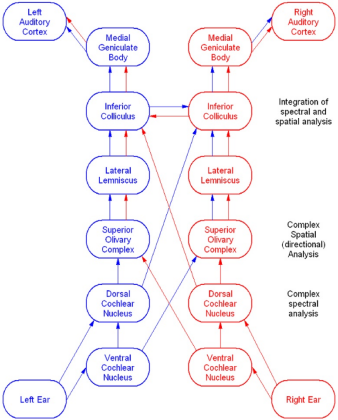
1. Why do robots need to hear?
2. Human-robot interaction
3. Auditory scene analysis
4. Audio signal processing in brief
5. Audio processing in the ear
6. **Audio processing in the midbrain**
7. Audio processing in the brain

Auditory Processing in the Brain

- The brain has the difficult task of transforming the acoustic-wave input into auditory object perception.
- The data flow travels from the ear to the auditory cortex via the midbrain that is composed of several *nuclei*
- The auditory cortex is divided into several areas, both its anatomy and organization varies considerably between the species.

The Ascending Auditory Pathway





Sound Localization in the Midbrain

- The superior olivary complex (SOC) is the first *station* that receives input from both ears
- The SOC is subdivided into two nuclei, the lateral superior olive (LSO) and the medial superior olive (MSO)
- These nuclei seem to be well suited to represent and measure time differences (MSO) and intensity differences (LSO)

Session Summary

- The midbrain is composed of nuclei
- These nuclei preprocess the audio information before is passed to the cortex
- Several nuclei receive input from both ears.

1. Introduction to Robot Hearing

1. Why do robots need to hear?
2. Human-robot interaction
3. Auditory scene analysis
4. Audio signal processing in brief
5. Audio processing in the ear
6. Audio processing in the midbrain
7. **Audio processing in the brain**

Auditory Cortex

The auditory cortex is largely understood today, there are two main hypotheses:

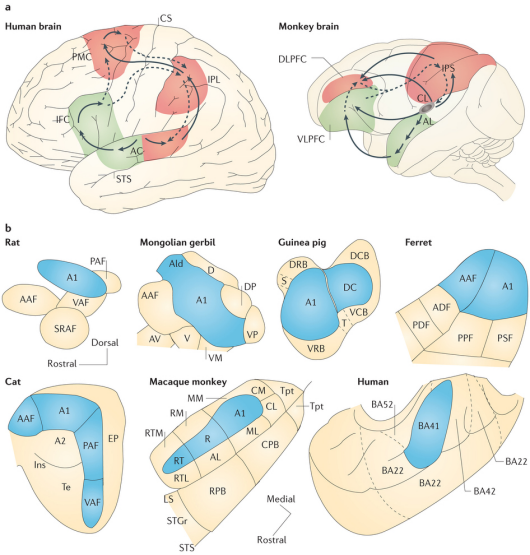
- There are two pathways, with a division of labor between spatial and non-spatial processing
- Dynamically organized processing networks are likely to support auditory perception.

The first hypothesis is supported by functional imaging (humans) and single-neuron physiological studies (non-human animals)

Auditory Streams in the Cortex

- Dorsal stream (red) analyze space and motion.
- Ventral stream (green) involved in auditory-object perception.
- Core regions (blue) of the auditory cortex for different species.

Auditory Areas



Nature Reviews | Neuroscience

Session Summary

- Division of labor: where and what pathways
- Dorsal and ventral streams
- The auditory cortex varies from species to species

Week Summary (I)

- It is necessary to augment physical interactions between robots and their environment with cognitive interactions between robots and people.
- Human-robot interaction (HRI) is more general and more difficult than human-computer interaction
- HRI goes well beyond speech processing/recognition, it needs auditory scene analysis
- Binaural hearing allows rich interactions based on acoustic wave processing and understanding

Week Summary (II)

- All animals (fish, frogs, reptiles, owls, bats, cats, monkeys, humans) have two ears...
- The anatomy/physiology of the ear is well understood... more than just a *microphone*.
- The midbrain plays a crucial role in spatial hearing, a simple physiological model for ITD/ILD processing is available.
- The auditory cortex processes in parallel two flow of information: where and what
- Biological models for the analysis of complex auditory information are not yet available.