Binaural Hearing for Robots

Methodological Foundations

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

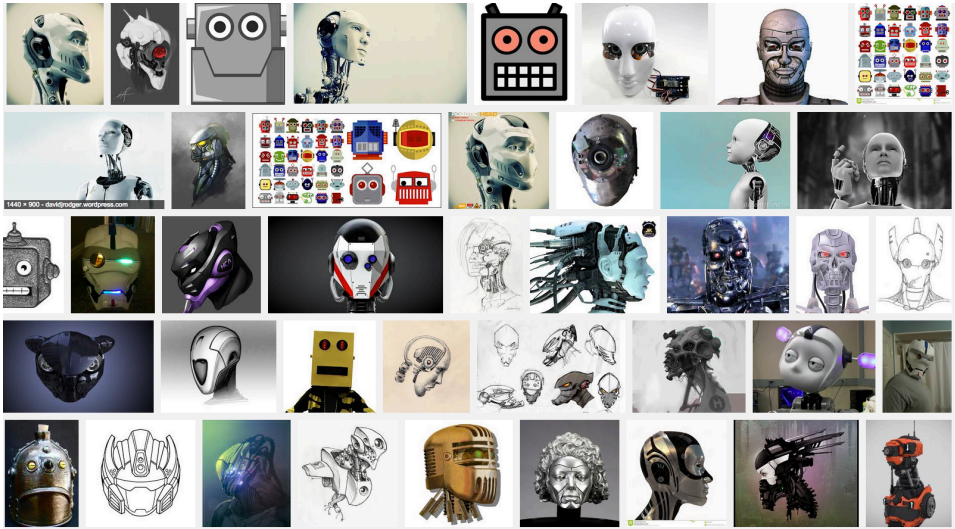
2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

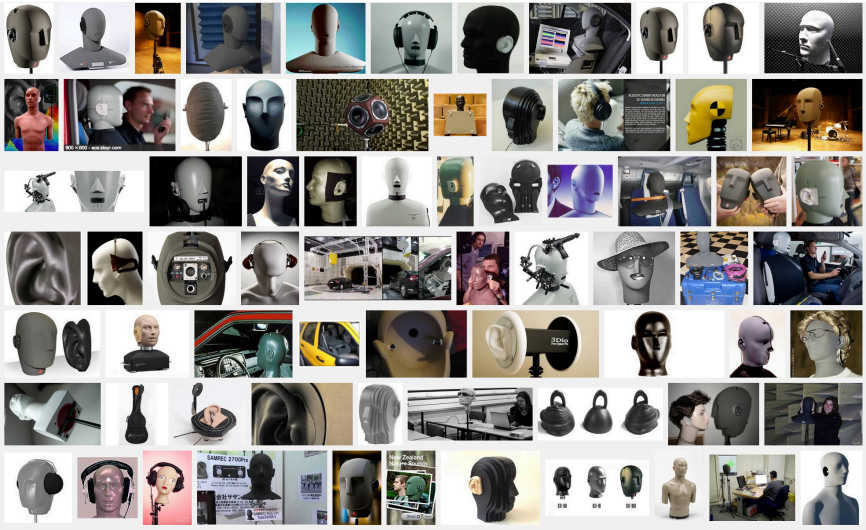
2. Methodological Foundations

1. **Robot heads and acoustic laboratories**
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Robot Heads

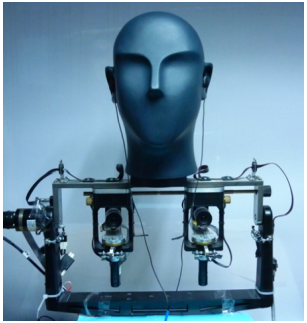


Acoustic Dummy Heads



Acoustic Head and Torso

Audio-Visual Robot Heads



Binaural & binocular heads mounted onto a pan-tilt device

Sound Sources

Sound sources: loudspeaker & people

An Acoustic Laboratory

Acoustic laboratory at Bar Ilan University

Anechoic Laboratories

Anechoic laboratories allow advanced acoustic experiments

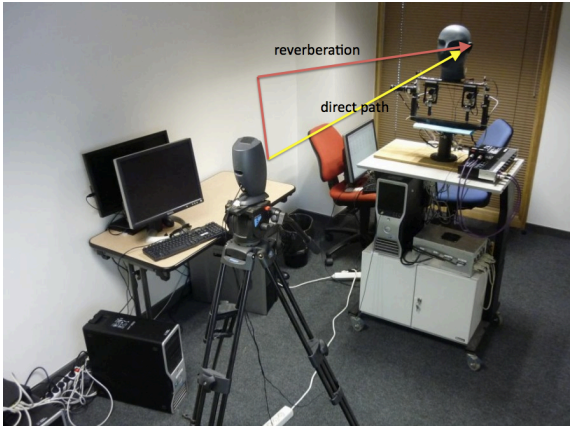
Session Summary

- Robot heads
- Acoustic heads
- Acoustic laboratories
- Anechoic rooms

2. Methodological Foundations

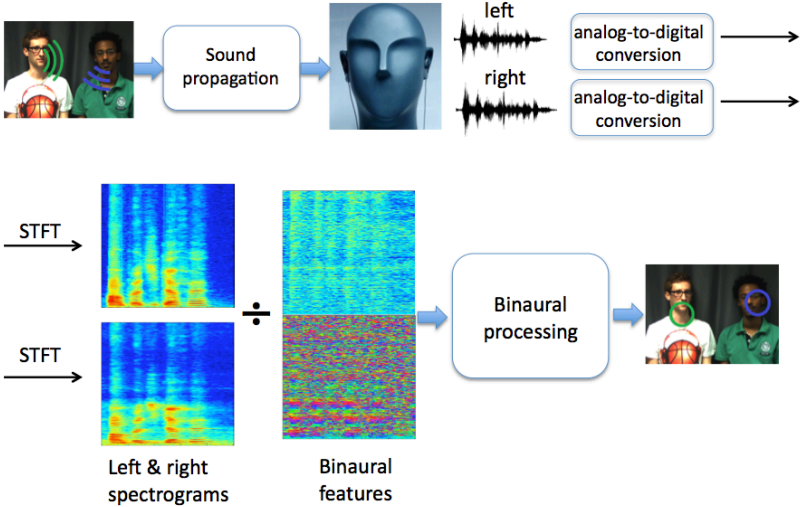
1. Robot heads and acoustic laboratories
2. **Binaural Processing Pipeline**
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

From the Source to the Microphones

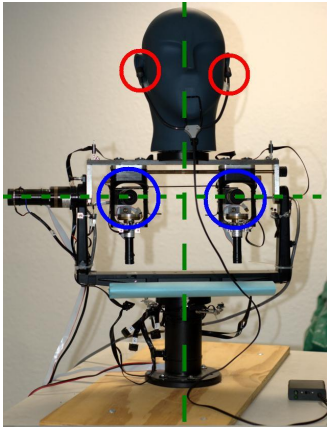


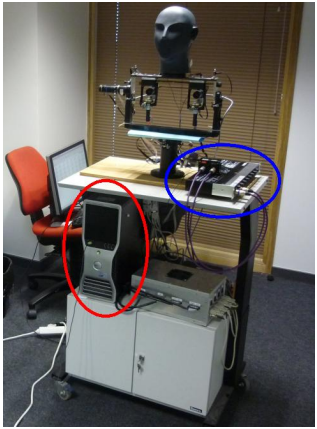
Standard robotic laboratory

Binaural Processing Pipeline



The Popeye Head





Fundamental Signal Processing

- Signal sampling (analog to digital conversion)
- Convolution (filtering)
- Correlation of two signals
- Fourier transforms
- Filtering and correlation in the Fourier domain

Session Summary

- Realistic laboratories
- The binaural pipeline
- Hardware components

2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. **Continuous-time Fourier transform**
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Continuous Source and Microphone Signals

- Signal emitted by a sound-source: $s(t) : \mathbb{R} \to \mathbb{R}$.
- Signal received by a microphone: $x(t) : \mathbb{R} \to \mathbb{R}$.
- In general $s(t)$ and $x(t)$ are **different**:
  The acoustic wave travels from source to microphone, hence the signal is modified.

Direct Sound Propagation Model

- Suppose that the wave travels from source to microphone along a straight line and that the wave **is not perturbed by room reverberations**:

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

- $n(t)$ is an additive noise.

More Realistic Sound Propagation Model

- When the signal (an acoustic wave) travels from source to microphone, it is **filtered** by an **unknown** **room impulse response** (RIR) denoted by $h(t)$:

$$x(t) = \int_{-\infty}^{+\infty} h(t-\tau)s(\tau)d\tau$$

- This operation is called **convolution**, denoted by "$\star$" and:

$$x(t) = h(t) \star s(t)$$

Continuous-time Fourier Transform

- A fundamental concept in signal processing is the Fourier transform (FT) of a continuous signal $x(t)$:

$$
\begin{aligned}
X(f) &= \int_{-\infty}^{+\infty} x(t) e^{-2\pi j f t} dt \\
&= \langle x, g^* \rangle, \quad \text{with} \quad g = e^{2\pi j f t}
\end{aligned}
$$

- $f \in \mathbb{R}$ is called the frequency and is measured in Hertz (Hz)
- $X(f) \in \mathbb{C}$: complex function that encodes both the amplitude and phase of a sinusoidal component of $x(t)$.

Convolution in the Spectral Domain

- Signal can be processed in the temporal or frequency (or spectral) domains.
- Temporal convolution becomes a product in the spectral domain. The equation $x(t) = h(t) \star s(t)$ becomes:

$$X(f) = H(f)S(f)$$

The Inverse Fourier Transform

- A signal can be **reconstructed** from its spectrum:

$$\tilde{x}(t) = \int_{-\infty}^{+\infty} X(f) e^{2\pi jft} df$$

- $\tilde{x}(t)$ is the reconstruction of $x(t)$ in the basis formed by the complex sinusoidal functions $\{e^{2\pi jft}\}_{f \in \mathbb{R}}$.

Session Summary

- Continuous signals
- Propagation models
- Fourier transform
- Convolution
- Inverse Fourier transform

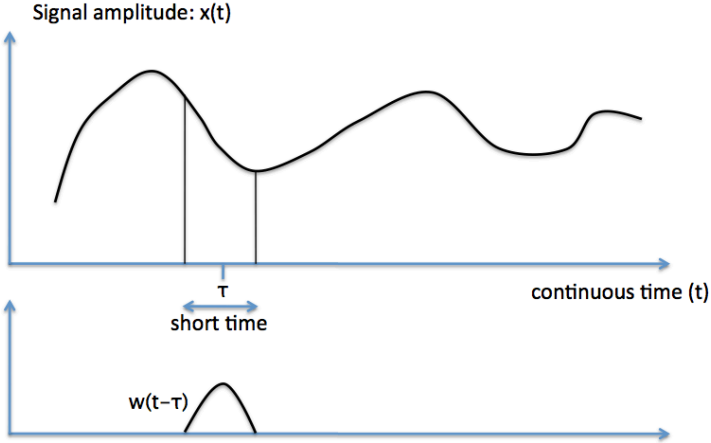
2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. **Continuous short-time Fourier transform**
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Short Time Fourier Transform (I)

- In practice, the Fourier transform is not very useful.
- The **short time Fourier transform** (STFT) is applied to the signal multiplied with a window function.
- This yields a two-dimensional representation or a *time-frequency* representation of the signal.

Windowing a Signal



Short Time Fourier Transform (II)

- The signal is multiplied with a **window function** which is non-zero for a short period of time:

$$X(f,\tau) = \int_{-\infty}^{+\infty} x(t)w(t-\tau)e^{-2\pi jft}dt$$

- The window function, centered at τ can be slid along the time axis.
- This yields a two-dimensional representation or a *time-frequency* representation of the signal.

Convolution with the STFT

- Using the short-time Fourier transform the temporal convolution

$$x(t) = h(t) \star s(t)$$

- becomes:

$$X(f, \tau) = H(f)S(f, \tau)$$

- The filter is constant over time, hence $H(f, \tau) = H(f)$.
- **Very important:**
  t (time index) is not the same as τ (slid index).

The Inverse STFT

- The signal can be recovered from the time-frequency representation:

$$\tilde{x}(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} X(f, \tau) e^{2\pi jft} d\tau df$$

Session Summary

- Windowing a signal
- STFT
- Convolution with the STFT
- Inverse STFT

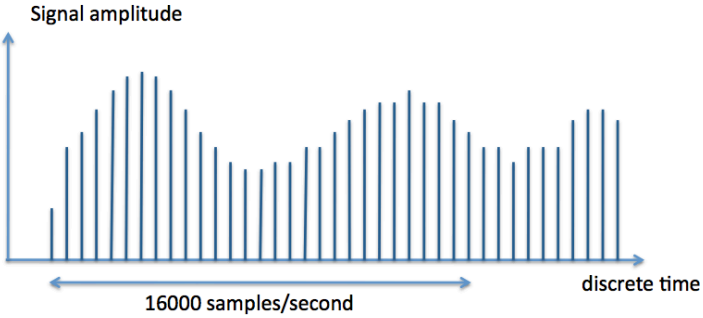
2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. **Discrete-time signals**
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Discrete-time Signals

- So far we considered continuous-time (analog) signals as they appear in nature.
- These signals must be sampled such that they can be processed by a computer.
- Sampling acoustic signals: $16,000$ **samples/second**.
- The amplitude is also quantized such that the signal becomes **digital**.

A Digital Signal



Notation

In many signal processing textbooks:

- The notation $x(t)$ is used for continuous signals, with $t \in \mathbb{R}$.
- The notation $x[n]$ is used for digital signals, with $n \in \mathbb{Z}$ (relative integer numbers).

For simplicity we will use the notation $x(t)$ for digital signals, with $t \in \mathbb{Z}$.

Discrete-time Fourier Transform (DFT)

- The Fourier transform can be written for digital signals:

$$X(f) = \frac{1}{N} \sum_{t=1}^{N} x(t) e^{-2\pi jft/N}$$

- f is an **integer**, for example $1, \ldots, f, \ldots, F$.
- $X(1), \ldots, X(f), \ldots, X(F)$ are complex numbers that encode the amplitude and phase of the signal for each frequency.

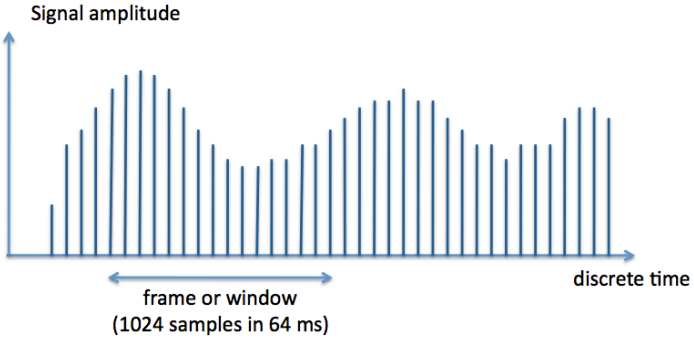
Session Summary

- Discrete signals
- Notation conventions
- Discrete Fourier transform

2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. **Discrete short-time Fourier transform**
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Digital Signal in a Temporal Window



Discrete-time STFT

- The Discrete-time Fourier transform is applied to a temporal window, or a **frame**:

$$X(f,\tau) = \frac{1}{N}\sum_{t=1}^{N} x(t)w(t-\tau)e^{-2\pi jft/N}$$

- τ is the frame index.
- The **frequency** f, **time** t, and **shift** τ are discrete.
- $X(f,\tau)$ is a complex-valued Fourier coefficient.

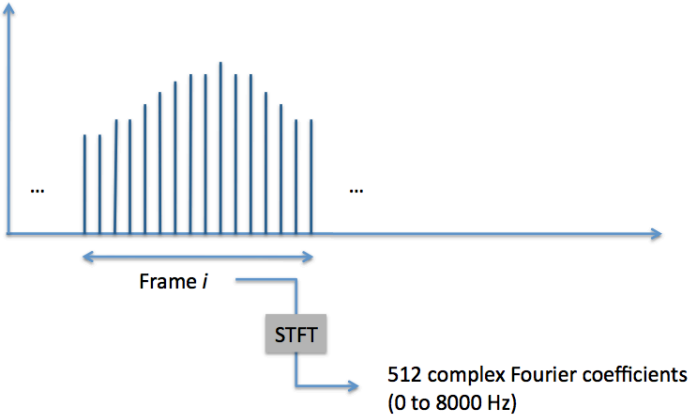
The Time-Shfit Theorem

- Let $X(f, \tau)$ be the STFT of $x(t)$.
- The STFT of x, **shifted** by t', is **circularly shifted** by a **phase**:

$$x(t - t') \Longrightarrow X(f, \tau) e^{2\pi j f t'/N}$$

- Notice that this phase shift $\alpha = 2\pi j f t'/N$ depends on the frequency f and on the time-shift t'.

The STFT for One Frame



Session Summary

- Windowing a discrete signal
- Discrete STFT
- Time-shift theorem
- The notion of frame

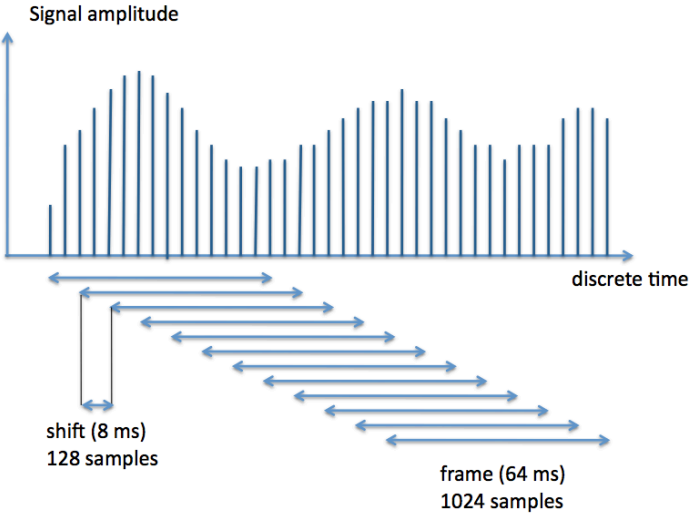
2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. **Spectrogram of an acoustic signal**
8. Cross-correlation
9. Relative transfer function
10. Binaural features

Building a Spectrogram

- We consider a frame of size 64 ms (or 1024 samples).
- The frame is slid along time with an 128 sample shift, (or 8 ms): $\tau_1 = 128$, $\tau_2 = 256$, $\ldots$, $\tau_8 = 1024$, etc.

Shifting the Frames Along the Signal

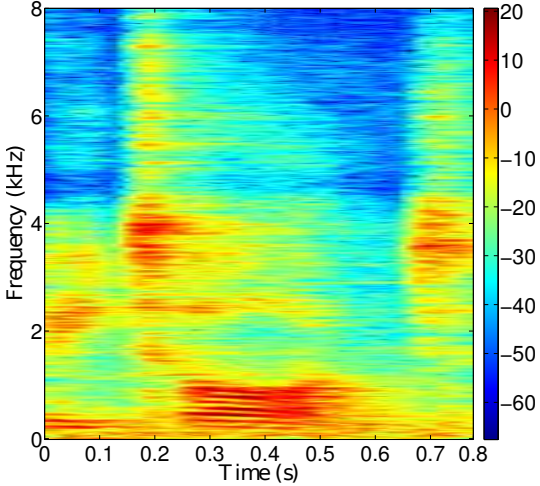


Spectrogram

In this way we can build a complex-valued spectrogram of a signal $x(t)$:

- $X(f, \tau) \in \mathbb{C}$ is also called a frequency-time coefficient with coordinates f and τ.
- $1 \leq f \leq F$ ($F = 512$ for a frequency range between 0 Hz and 8000 Hz).
- $1 \leq \tau \leq L$ (there are 126 frames for a signal of one second)
- Spectrogram visualization, the module: $\mathcal{X}(f, \tau) = |X(f, \tau)|$.

The Spectrogram of a Speech Signal



Session Summary

- How to build a spectrogram
- Spectrogram structure
- Spectrogram of speech

2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. **Cross-correlation**
9. Relative transfer function
10. Binaural features

Cross-Correlation of Two Signals

- The cross-correlation (CC) function is a fundamental tool for comparing two signals:

$$r_{x_1,x_2}(\tau) = \sum_{t=1}^{T} x_1(t)x_2(t-\tau)$$

- Cross-correlation is a fundamental tool in binaural hearing as it allows to estimate the time difference between two signals recorded with two microphones:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmax}}\, r_{x_1,x_2}(\tau)$$

Normalized Cross-Correlation

- The normalized cross-correlation (NCC) function varies between 0 (fully uncorrelated signals) and 1 (identical signals):

$$\rho_{x_1,x_2}(\tau) = \frac{r_{x_1,x_2}(\tau)}{\left(r_{x_1,x_1}(0) r_{x_2,x_2}(0)\right)^{1/2}}$$

- $r_{x,x}(0) = \sum_{t=1}^{T} |x(t)|^2$ is also called the power of the signal.

The Power Spectral Density (PSD)

- In the spectral domain, one can compute the signal power for each frequency.
- First, we compute the power over a frame l

$$\phi_{x,x}(f,l) = X(f,l)X^*(f,l) = |X(f,l)|^2$$

- Second we compute the average power over several frames:

$$\Phi_{x,x}(f) = \frac{1}{L}\sum_{l=1}^{L}\phi_{x,x}(f,l)$$

The Cross-Power Spectral Density

- We can also define the **cross-power** of two signals for a frame l:

$$\phi_{x_1,x_2}(f,l) = X_1(f,l)X_2^*(f,l),$$

- as well as the average power over several frames:

$$\Phi_{x_1,x_2}(f) = \frac{1}{L}\sum_{l=1}^{L}\phi_{x_1,x_2}(f,l)$$

Session Summary

- Correlating two signals
- Power spectral density
- Cross-correlation in the spectral domain

2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. **Relative transfer function**
10. Binaural features

Using Two Microphones

- We now address the problem of defining some kind of **binaural features** that can be used to characterize sounds and to extract source parameters.
- Propagation model:

$$x_1(t) = h_1(t) \star s(t)$$
$$x_2(t - t') = h_2(t) \star s(t)$$

- The signal x_2 is delayed by t' with respect to x_1.
- In the spectral domain we have:

$$X_1(f, l) = H_1(f)S(f, l)$$
$$X_2(f, l)e^{\alpha(f,l,t')} = H_2(f)S(f, l)$$

Binaural Transfer Function

- The **relative transfer function** (RTF):

$$H_{\mathrm{rel}}(f) = \frac{H_2(f)}{H_1(f)}$$

- The RTF encodes binaural information:

$$H_{\mathrm{rel}}(f) \approx \frac{|X_2(f,l)|}{|X_1(f,l)|} e^{\alpha(f,l,t')}$$

Relative Transfer Function

- The RTF is a complex-valued function.
- The magnitude of the RTF encodes the ratio of the left and right magnitudes for each frequency f and frame l
- The argument of the RTF depends on the time-delay between the two microphone signals:

$$H_{\mathrm{rel}}(f, l) = \mathcal{H}_{\mathrm{rel}}(f, l) e^{\alpha(f, l, t')}$$

Estimation of the RTF

- We consider the power spectral density and cross-power spectral density of the two microphone signals:

$$\Phi_{x_1,x_1}(f) = \frac{1}{L} H_1(f) H_1^*(f) \sum_{l=1}^{L} |S(f,l)|^2$$

$$\Phi_{x_1,x_2}(f) = \frac{1}{L} H_1(f) H_2^*(f) \sum_{l=1}^{L} |S(f,l)|^2$$

- The RTF can be estimated with:

$$\hat{H}_{\mathrm{rel}}(f) \approx \frac{\Phi_{x_1,x_2}(f)}{\Phi_{x_1,x_1}(f)}$$

A Robot Head

Binaural head

Sound sources

The Head Related Transfer Function

- If there are no reverberations (anechoic environment) the impulse responses h_1 and h_2 model the filtering effects of the head and of the pinnae.
- In this case, the RTF is called the **head related transfer function** (HRTF):

$$H_{\text{head}}(f) = \frac{H_2(f)}{H_1(f)}$$

- The HRTF can be estimated using the same principle as the RTF, for each frame l:

$$\hat{H}_{\text{head}}(f, l) \approx \frac{\phi_{x_1,x_2}(f, l)}{\phi_{x_1,x_1}(f, l)}$$

Session Summary

- Relative transfer function
- Estimation of RTF
- Back to a robot head
- Head-related transfer function

2. Methodological Foundations

1. Robot heads and acoustic laboratories
2. Binaural Processing Pipeline
3. Continuous-time Fourier transform
4. Continuous short-time Fourier transform
5. Discrete-time signals
6. Discrete short-time Fourier transform
7. Spectrogram of an acoustic signal
8. Cross-correlation
9. Relative transfer function
10. **Binaural features**

Binaural Cues

- The log-magnitude of HRTF is called the **interaural level difference** (ILD);
- The argument of HRTF is called the **interaural phase difference** (IPD):

$$\mathrm{ILD}(f, l) = 20 \log |\hat{H}_{\mathrm{head}}(f, l)|$$
$$\mathrm{IPD}(f, l) = \arg(\hat{H}_{\mathrm{head}}(f, l))$$

- The IPD is the equivalent of a time-delay, **interaural time difference** (ITD).

ILD Vector

- The ILD is the magnitude of the head-related transfer function:

$$\mathrm{ILD}(f, l) \in \mathbb{R}, \forall f, 1 \leq f \leq F, \forall l, 1 \leq l \leq L$$

- The ILD values corresponding to the l-frame form a vector of dimension F:

$$\mathbf{ILD}(l) = \left(\mathrm{ILD}(1, l), \ldots, \mathrm{ILD}(f, l), \ldots, \mathrm{ILD}(F, l)\right)^\top \in \mathbb{R}^F$$

IPD Vector

- The IPD is the phase of the head-related transfer function, and it can be represented by its real and imaginary parts:

$$\mathrm{ILD}(f, l) \in \mathbb{R}^2, \forall f, 1 \leq f \leq F, \forall l, 1 \leq l \leq L$$

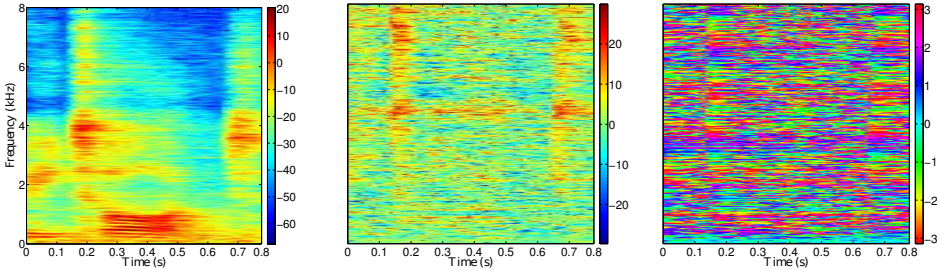
- The IPD values corresponding to the l-frame form a vector of dimension $2F$:

$$\mathbf{IPD}(l) = \big(\mathrm{IPD}(1, l), \ldots, \mathrm{IPD}(f, l), \ldots, \mathrm{IPD}(F, l)\big)^\top \in \mathbb{R}^{2F}$$

Binaural Spectrograms

- The ILD and IPD vectors can form spectrograms.
- A spectrogram can be viewed as a time-series of L vectors:
- **ILD spectrogram**: $\mathbf{ILD}(1), \ldots, \mathbf{ILD}(l), \ldots, \mathbf{ILD}(L)$
- **IPD spectrogram**: $\mathbf{IPD}(1), \ldots, \mathbf{IPD}(l), \ldots, \mathbf{IPD}(L)$

ILD and IPD Spectrograms of a Speech Signal

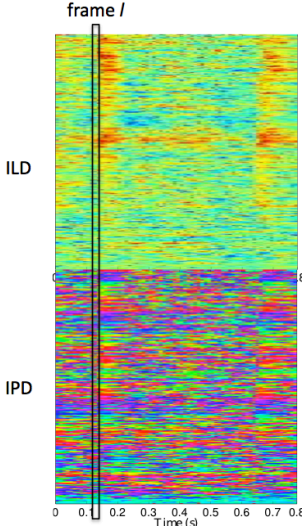


Speech (left microphone) ILD IPD

ILPD Vectors

- Concatenate ILD and IPD vectors for each frame l.
- We obtain **ILPD vectors** $\boldsymbol{y}(1), \ldots, \boldsymbol{y}(l), \ldots, \boldsymbol{y}(L) \in \mathbb{R}^{3F}$.
- Example: $F = 512$ (number of frequencies used to represent the signal in the Fourier domain)
- $\boldsymbol{y}(l) \in \mathbb{R}^{1536}$ is a **high-dimensional** vector.

Example of an ILPD Vector



How To Compute ILPD Vectors?

- Record a sound of length T with two microphones plugged in a robot head and discretize the signals at **16000 samples per second**, $x_1(t), x_2(t)$.
- Compute the STFT representations of the two signals, L frames with **128 samples per frame**, $\boldsymbol{X}_1(l), \boldsymbol{X}_2(l)$.
- Estimate the ILD and IPD for each frame.
- Concatenate the ILD and IPD.

Session Summary

- Binaural cues
- Interaural level difference (ILD)
- Interaural phase difference (IPD)
- Binaural spectrograms
- ILPD features

Week Summary

- Robot heads with ears, sound sources with known properties, acoustic laboratories.
- Binaural processing pipeline: from speakers to their locations and identity.
- Signal processing in the continuous and discrete domains.
- Spectrograms, transfer functions, binaural features.