Binaural Hearing for Robots

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

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
The ILPD representation

We already introduced the ILPD representation:

- The ILD and IPD are computed from the estimated HRTF (head related transfer function):

\[
\text{ILD}(f, l) = 20 \log |\hat{H}_{\text{head}}(f, l)| \in \mathbb{R}
\]

\[
\alpha_{f,l} = \arg(\hat{H}_{\text{head}}(f, l))
\]

\[
\text{IPD}(f, l) = (\cos \alpha_{f,l}, \sin \alpha_{f,l}) \in \mathbb{R}^2
\]

- \(\text{ILPD}(f, l) \in \mathbb{R}^3\) is the concatenation of ILD and IPD.
ILPD for one Frame

- For each frame $l, 1 \leq l \leq L$, the ILPD vector is of dimension $3F$.
- Call this vector $y_l$:

$$y_l = \left( \text{ILD}(1, l), \cos \alpha_{1,l}, \sin \alpha_{1,l}, \right.$$  
$$\ldots, \text{ILD}(f, l), \cos \alpha_{f,l}, \sin \alpha_{f,l},$$  
$$\ldots, \text{ILD}(F, l), \cos \alpha_{F,l}, \sin \alpha_{F,l} \right)^T$$

- For $F = 512$, $y \in \mathbb{R}^{1536}$;
- This is a high-dimensional vector space!
An ILPD spectrogram with $L$ frames is a time-series of ILPD vectors, or a matrix:

$$Y = (y_1 \ldots y_i \ldots y_L) \in \mathbb{R}^{3F \times L}$$
Sound Types

- Sparse-band and narrow-band signals: in natural sounds only a few frequency are significant, the other frequencies are missing,
- Broad-band signals: in white-noise sounds all the frequencies are significant.
Examples

Speech

White noise
Binary Masking

• To fully characterize a spectrogram, we introduce a \textbf{binary-mask} matrix $\Lambda$:

$$\Lambda_{f,l} = \begin{cases} 
1 & \text{if } |X_1(f, l)|^2 + |X_2(f, l)|^2 \geq -20\text{dB} \\
0 & \text{otherwise}
\end{cases}$$

• $|X_1(f, l)|^2 + |X_2(f, l)|^2$ is the \textbf{total power spectral density} (total-PSD) of the binaural recordings.
A Sparse ILPD Spectrogram

\[ \Lambda_{f,i} = 0 \]
A spectrogram $S$ is a matrix of dimensions $3F \times L$, also a time-series of $L$ ILPD vectors, and a binary-mask matrix:

$$S = \{(y_1, \ldots, y_i, \ldots, y_L), \Lambda\} = \{Y, \Lambda\}$$
Session Summary

- Putting ILD and IPD features together
- ILPD spectrogram
- Broad-band and narrow-band sounds
- Sparse spectrograms
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. **Mapping Sounds onto Their Directions**
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Binaural Features and Source Direction

- Remember the link between IPD (interaural phase difference) and TDOA (time difference of arrival).
- This link is a direct consequence of the time-shift theorem (week #2).
- Also, there is a link between TDOA and source direction in case of direct propagation.
- In the presence of filtering effects, such as the HRTF (head-related transfer function) there is NO explicit relationship.
ILPD and Sound-Source Direction

- Let $y \in \mathbb{R}^D$ be an ILPD vector observed with a binaural robot head.
- $D = 3F$, or the concatenation of ILD (of size $F$) and IPD (of size $2F$), for example $F = 512$ frequencies.
- Let $x \in \mathbb{R}^2$ be the direction (azimuth and elevation) of a sound source.
Acoustic to Direction Mapping

- We seek an **explicit** representation of the mapping:

  \[ \text{ILPD vector} \rightarrow \text{source direction} \]

- Or a function \( f \) that maps **high-dimensional** ILPD vectors onto **low-dimensional** directions:

  \[ x = f(y), \quad f : \mathbb{R}^D \rightarrow \mathbb{R}^2 \]
Unsupervised learning: Both $f$ and $x$ are unknown.

1. Sample the high-dimensional space, $y_1, \ldots, y_n, \ldots, y_N$;
2. Extract a low-dimensional manifold from the high-dimensional sample space;

- $f$ in $x = f(y)$ is a **linear** or **non-linear** projection of $\mathbb{R}^D$ onto $\mathbb{R}^2$.
- This is referred to as:
- **dimension reduction** (PCA) or
- **manifold learning** (LE, LTSA).
Learning an Acoustic-to-Direction Mapping (II)

Supervised learning: Only \( f \) is unknown.

1. Sample the high-dimensional space, \( y_1, \ldots, y_n, \ldots, y_N \);
2. Observe sound directions, \( x_1, \ldots, x_n, \ldots, x_N \);
3. Form an input-output set of training samples, 
   \((y_1, x_1), \ldots, (y_n, x_n), \ldots, (y_N, y_N)\)

   • Estimate \( f \) from:

   \[
   \begin{align*}
   x_1 &= f(y_1), \\
   \vdots \\
   x_n &= f(y_n), \\
   \vdots \\
   x_N &= f(y_N).
   \end{align*}
   \]

   • This is referred to as regression.
Regression

- **Linear regression**, a projection of $\mathbb{R}^D$ onto $\mathbb{R}^2$:

  $$x = Ay + b, \quad A \in \mathbb{R}^{2 \times D}, \quad b \in \mathbb{R}^2$$

- **Piecewise-linear regression** (there are $K$ possible projections):

  $$x = \sum_{k=1}^{K} \mathbb{I}(z = k)(A_k y + b_k)$$

- $\mathbb{I}(z)$ is called an *indicator function*, that selects the $k$-th affine transformation $A_k, b_k$.

- We will build a binaural localization method based on piecewise-linear regression.
Session Summary

- Link between binaural features and sound localization
- Learning an acoustic-to-direction mapping
- Unsupervised learning
- Supervised learning
- Regression
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. **Collecting Training Data**
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Principle of Data Collection

Binaural recording:
- Right
- Left

Extract auditory feature vector

Source emitting from position $x_n$

Repeat for $N$ positions

Extracted auditory feature vector

$y_n \in \mathbb{R}^D$

Sampled acoustic space

$\{y_n\}_{n=1}^{N}$
Datasets

- **Training data**: White noise (broad-band sounds) emitted by a loudspeaker in known directions.
- **Test data**: Speech (sparse-band sounds) emitted by a loudspeaker and by people.
- We use two configurations:
  1. **Audio-motor data**: The robot head rotates while the sound source is fixed.
  2. **Audio-visual data**: The loudspeaker is moved while the robot head is fixed.
Setup of Audio-Motor Data Collection
Example of Audio-Motor Data Collection

Insert AMtraining.wmv here
Setup of Audio-Visual Data Collection
Sound Direction from an Image

direction: \((\alpha, \beta)\)

image

pixel: \((i,j)\)

focal center

sound source
Example of Audio-Visual Data Collection

Insert av_calibration.wmv here
Session Summary

- Collecting data with ground-truth
- Audio-motor data collection
- Audio-visual data collection
- Practical setup
- Examples of data collection campaigns
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method

Radu Horaud
Binaural Hearing for Robots
Localization on the Binaural Manifold

Unsupervised learning: Both $f$ and $x$ are unknown.

1. Sample the high-dimensional space, $y_1, \ldots, y_n, \ldots, y_N$;
2. Extract a low-dimensional manifold from the high-dimensional sample space;

- $f$ in $x = f(y)$ is a **linear** or **non-linear** projection of $\mathbb{R}^D$ onto $\mathbb{R}^2$.
- This is referred to as:
  - **dimension reduction** (PCA) or
  - **manifold learning** (LE, LTSA).
Manifold Learning

- Extract a low-dimensional representation from high-dimensional data.
- Examples: principal component analysis (PCA), Laplacian embedding (LE), local tangent space analysis (LTSA), etc.
- These methods find a linear subspace (PCA), non-linear subspace (LE) or piecewise linear subspace (LTSA).
The Task of Manifold Learning

High-dimensional space of binaural features (1536)

Low-dimensional subspace of sound directions (2)
Examples of Manifold Learning

Principal component analysis  Local tangent-space analysis
Session Summary

- Analyzing binaural features
- The binaural manifold
- Dimensionality reduction
- Principal component analysis
- Local tangent space analysis
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. **Localization with a Look-up Table**
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Localization Based on a Look-up Table

• Consider a training dataset $\mathcal{T} = \{(y_1, x_1), \ldots, (y_n, x_n), \ldots, (y_N, y_N)\}$.
• Each ILPD vector $y_n \in \mathbb{R}^D$ corresponds to white-noise recordings of length 1 s.
• There is a sound-source direction $x_n \in \mathbb{R}^2$ associated with each ILPD white-noise vector.
• We want to localize two types of sound-sources:
  1. Broad-band (white-noise) sounds, with ILPD vectors $w \in \mathbb{R}^D$.
  2. ILPD speech spectrograms, denoted $S = \{s_1, \ldots, s_l, \ldots, s_L, \Lambda\}$
• In this spectrogram $s_l \in \mathbb{R}^D$ and $\Lambda$ is a binary mask.
Localization with Nearest Neighbor Search (Noise)

- A white-noise emitter can be localized in a straightforward manner using nearest-neighbor search:

\[ \hat{n} = \arg\max_n \| \mathbf{w} - \mathbf{y}_n \|^2 \]

- The sound direction is simply:

\[ \hat{x} = x_{\hat{n}} \]
Localization with Nearest Neighbor Search (Speech)

- This can be extended to localize sounds with missing frequencies (speech):

\[ \hat{n} = \arg\max_n \sum_{l=1}^{L} \sum_{f=1}^{D} \Lambda_{fl}(s_{fl} - y_{fn})^2 \]

- The sound direction is simply:

\[ \hat{x} = x_{\hat{n}} \]
Probabilistic Framework (I)

- We introduce a **generative model**, namely $D$ functions $g_f$ that map the source direction onto an ILPD value:

$$g_f : \mathbf{x} \rightarrow y_f, \forall f, 1 \leq f \leq D.$$  

- A frequency-time point of a spectrogram $S$ is a **Gaussian random variable** with probability distribution function (pdf) given by:

$$p(s_{fl}; \mathbf{x}, \sigma_f) = \frac{1}{\sigma_f \sqrt{2\pi}} \exp \left( - \frac{(s_{fl} - g_f(\mathbf{x}))^2}{2\sigma_f^2} \right)$$

- where the **mean** and **variance** are $g_f(\mathbf{x})$ and $\sigma_f^2$, respectively
Probabilistic Framework (II)

- The spectrogram points are independent and identically distributed (iid), or
  \[ p(s_{11}, \ldots, s_{f_l}, \ldots, s_{DL}) = \prod_{f=1}^{D} \prod_{l=1}^{L} p(s_{fl}). \]

- The log-likelihood is:
  \[ L(S; \mathbf{x}, \sigma_f) = \sum_{f=1}^{D} \sum_{l=1}^{L} p(s_{fl}; \mathbf{x}, \sigma_f) \]

- Sound direction is obtained by maximization of this function:
  \[ \hat{x} = \arg\max_x L(S; x, \sigma_f) \]

- This can be achieved if the mappings \( g_f \) are learnt from the training set.
Session Summary

- Non-parametric localization
- Localizing white-noise
- Localizing speech
- Introduction to probabilistic setting
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. **Linear Regression**
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Multivariate Linear Regression (I)

- Consider the regression problem already described and applied to the training set: $\mathcal{T} = \{(y_1, x_1), \ldots, (y_n, x_n), \ldots, (y_N, y_N)\}$:
  1. Estimate $\hat{f}$ from:

\[
x_1 = f(y_1), \\
\vdots \\
x_N = f(y_N).
\]

  2. Predict the direction of a white-noise emitter:

\[
\hat{x} = \hat{f}(w)
\]
Multivariate Linear Regression (II)

- Estimate matrix $A \in \mathbb{R}^{2 \times D}$ and vector $b \in \mathbb{R}^{2}$ (affine transformation):

$$x_1 = Ay_1 + b,$$

$$\vdots$$

$$x_n = Ay_n + b,$$

$$\vdots$$

$$x_N = Ay_N + b.$$
Linear Regression Formulation

• These equations can be rearranged to yield a matrix-vector equation of the form:

\[ Y a = X \]

• with:
  1. \( Y \in \mathbb{R}^{(2N) \times (2D+2)} \) is a matrix containing the observed input training data (ILPD vectors).
  2. \( a \in \mathbb{R}^{2D+2} \) is a vector containing the unknown entries of \( A \) and \( b \).
  3. \( X \in \mathbb{R}^{2N} \) is a vector containing the observed output training data (sound directions).
• Each input-output pair \((y_n, x_n)\) yields 2 linear constraints.
• A solution is possible if \(Y \in \mathbb{R}^{(2N) \times (2D+2)}\) is invertible, or if \(N \geq D + 1\).
• The dimension of the ILPD vectors: \(D = 1536\),
• A solution exists only if \(N \geq D + 1 = 1537\) training pairs are available:

\[
\hat{a} = (Y^T Y)^{-1} Y^T X
\]
Interchanging Input and Output

- Linear regression may also be estimated **the other way around**: Estimate matrix \( A \in \mathbb{R}^{D \times 2} \) and vector \( b \in \mathbb{R}^{D} \):

\[
\begin{align*}
y_1 &= Ax_1 + b, \\
\vdots \\
y_n &= Ax_n + b, \\
\vdots \\
y_N &= Ax_N + b.
\end{align*}
\]
Reversed Linear Regression

\[ Xa = Y \]

with:

1. \( X \in \mathbb{R}^{(DN) \times (2D+D)} \) is a matrix containing the observed output training data (sound directions).
2. \( a \in \mathbb{R}^{2D+D} \) is a vector containing the unknown entries of \( A \) and \( b \).
3. \( Y \in \mathbb{R}^{DN} \) is a vector containing the observed input training data (ILPD vectors).
Reversed Least-Square Solution

- Each output-input pair \((x_n, y_n)\) yields \(D\) linear constraints.
- A solution is possible if \(Y \in \mathbb{R}^{(DN) \times (2D+D)}\) is invertible: \(N \geq L + 1 = 3\).
- A minimum of \(N = 3\) pairs are needed.
- With \(N \gg 3\), the system is over constrained:

\[
\hat{a} = (X^T X)^{-1} X^T Y
\]
Session Summary

- Multivariate linear regression
- Computational complexity
- Interchanging the input and the output
- Least-square solution
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. **Piecewise Linear Regression**
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Regression

- Reversed Linear regression, \(x \in \mathbb{R}^2, y \in \mathbb{R}^D:\)
  \[
y = Ax + b, \quad A \in \mathbb{R}^{D \times 2}, b \in \mathbb{R}^D
  \]

- **Piecewise linear regression** (there are \(K\) possible mappings):
  \[
y = \sum_{k=1}^{K} \mathbb{I}(z = k)(A_k x + b_k + e_k)
  \]
  - \(\mathbb{I}(z)\) is called an **indicator function**, that selects the \(k\)-th affine transformation \(A_k, b_k\).
  - \(e_k \in \mathbb{R}^D\) is an error vector accounting for the piecewise linear linear approximation.
Probabilistic Setting

• The joint input-output probability is decomposed:

\[ p(y, x) = \sum_{k=1}^{K} p(y|x, Z = k)p(x|Z = k)p(Z = k) \]

• Assuming Gaussian (normal) distributions we have:

\[ p(y|x, Z = k) = \mathcal{N}(y; A_kx + b_k, \Sigma) \]
\[ p(x|Z = k) = \mathcal{N}(x; c_k, \Gamma_k) \]
\[ p(Z = k) = \pi_k \]
Gaussian Mixture Model

- This formulation belongs to Gaussian Mixture Models (GMM).
- The model parameters are:

\[
\theta = \{c_k, \Gamma_k, \pi_k, A_k, b_k, \Sigma_k\}_{k=1}^K
\]

- The parameters can be estimated via an expectation-maximization (EM) algorithm:
  1. Initialize the model parameters \(\theta^{(0)} = \{c_k^{(0)}, \Gamma_k^{(0)}, \pi_k^{(0)}, A_k^{(0)}, b_k^{(0)}, \Sigma_k^{(0)}\}_{k=1}^K\),
  2. Evaluate the posterior probabilities \(r_{kn}^{(i)} = p(Z_n = k|x_n, y_n; \theta^{(i-1)})\),
  3. Maximize the complete-data expected log-likelihood:

\[
\theta^{(i)} = \arg \max_{\theta} \mathbb{E}[ \log p(X, Y, Z|\theta; \theta^{(i-1)})].
\]
  4. Iterate steps 2 & 3 until convergence.
Posterior Probabilities (I)

- The optimal parameters $\hat{\theta}$ thus obtained allow to estimate the probability of $y$ given $x$:

$$p(y|x; \hat{\theta}) = \sum_{k=1}^{K} \frac{\pi_k \mathcal{N}(x; \hat{c}_k, \hat{\Gamma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x; \hat{c}_j, \hat{\Gamma}_j)} \mathcal{N}(y; \hat{A}_k x + \hat{b}_k, \hat{\Sigma}_k).$$
Posterior Probabilities (II)

• More interesting, one can also evaluate the posterior probability of the sound direction, \( x \), given an ILPD vector, \( y \):

\[
p(x|y; \tilde{\theta}) = \sum_{k=1}^{K} \frac{\bar{\pi}_k \mathcal{N}(y; \tilde{c}_k, \tilde{\Gamma}_k)}{\sum_{j=1}^{K} \bar{\pi}_j \mathcal{N}(y; \tilde{c}_j, \tilde{\Gamma}_j)} \mathcal{N}(x; \tilde{\mu}_k, \tilde{\Sigma}_k)
\]

• The parameters \( \tilde{\theta} \) are closed-form expressions of \( \hat{\theta} \). \(^1\)

\(^1\)https://hal.inria.fr/hal-00863468/en
Session Summary

- Probabilistic treatment of regression
- Gaussian mixture model for regression
- Parameter estimation
- Bayes inversion
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
The regression just described, once learned using a training dataset\n\[ \mathcal{T} = \{ y_n, x_n \}_{n=1}^{N} \], can be used to predict the direction \( x \) of a sound.

There are two cases:

1. Broad-band (white-noise) sounds, with ILPD vectors denoted \( w \in \mathbb{R}^D \).
2. ILPD speech spectrograms (sparse), denoted \( S = \{ s_1, \ldots, s_l, \ldots, s_L, \lambda \} \).
Localizing Broad-Band Sounds

- Recall the posterior of a sound direction given a broad-band ILPD vector $w$, and the regression parameters:

$$
p(x|w; \tilde{\theta}) = \sum_{k=1}^{K} \tilde{\nu}_k \mathcal{N}(x; \tilde{\mu}_k, \tilde{\Sigma}_k)
$$

- The optimal direction of a broad-band sound (all the frequencies are active) can be evaluated with:

$$
\tilde{x} = \mathbb{E}[x|w; \tilde{\theta}] = \sum_{k=1}^{K} \tilde{\nu}_k \tilde{\mu}_k
$$
Localizing Speech

• Speech is described by a sparse ILPD spectrogram \( S = \{s_1, \ldots, s_l, \ldots, s_L, \Lambda\} \)
• There is an equivalent posterior for sparse-spectrum sounds:

\[
p(x|S; \bar{\theta}) = \sum_{k=1}^{K} \tilde{\nu}_k(S, \bar{\theta}) \mathcal{N}(x; \tilde{\mu}_k(S, \bar{\theta}), \tilde{S}_k(S, \bar{\theta}))
\]

• In particular, the GMM parameters (proportions, means, and covariances) depend on the binary mask \( \Lambda \) of the speech spectrogram.
• The optimal direction of speech can be evaluated with:

\[
\tilde{x} = \mathbb{E}[x|S; \bar{\theta}] = \sum_{k=1}^{K} \tilde{\nu}_k \tilde{\mu}_k
\]

\(^2\)https://hal.inria.fr/view/index/docid/1112834
Data Collection Scenario

Test scenario (left) & Training (right)

Live experiments (speech turns)
Horizontal (azimuth) direction is more precise than vertical (elevation) direction.
Session Summary

- Sound localization based on a learned regression
- Training done with broad-band sounds
- Localization of sparse-band sounds is possible
- Probabilistic setting
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. Separation & Localization Method
Sound Separation Pipeline

What’s your name?

Hi, my name is Israel
Spectrograms of Speech Mixtures

[Diagram showing the process of binaural processing with audio samples and spectrograms]
Separation with Binary Masking (I)
Separation with Binary Masking (II)

**HOW TO COMPUTE THE MASKS?**

- STFT
- iSTFT
- Vincent
- Output
- Israel
Outline of Methodology

- Learning binaural-to-source mappings
  1. Collect a training dataset \( \mathcal{T} = \{ y_n, x_n \}_{n=1}^{N} \) corresponding to a **single broad-band source** recorded with a binaural head.
  2. Estimate the model parameters \( \tilde{\theta} \) that allow to evaluate the posterior probability of an unknown source direction from an observed spectrogram.

- Simultaneous separation and localization
  1. Consider a recorded **binaural spectrogram** \( S \) that contains a mixture of \( M \) sound sources with **unknown directions** and with **unknown binary masks**.
  2. A separation and localization method must:
     - Assign a binary mask to each source \( m \), and
     - Estimate a direction \( \tilde{x}_m \) for each source \( m \).
Session Summary

- Principles of sound-source separation
- Mixed spectrogram
- Binary masking
- Separation and localization methodology
4. Machine Learning and Binaural Hearing

1. Binaural Features
2. Mapping Sounds onto Their Directions
3. Collecting Training Data
4. The Binaural Manifold
5. Localization with a Look-up Table
6. Linear Regression
7. Piecewise Linear Regression
8. Predicting the Direction of Speech
9. Principles of Sound Separation
10. *Separation & Localization Method*
Sound Separation and Localization

- We describe a method that combines separation and localization.
- **Separation** uses the concept of binary masking: Starting from the binaural spectrogram $S$, which contains a mixture of all the emitting sources, a binary mask $\chi_m$ for each source is estimated. These masks are used to split $S$ into $M$ disjoint spectrograms $S_1, \ldots, S_m, \ldots, S_M$.
- **Localization** The same source-direction principle as before is applied to each spectrogram $S_m$ and associated binary mask $\chi_m$ to estimate the direction $\tilde{\mathbf{x}}_m$ of source $m$.
- A variational expectation-maximization (VEM) algorithm implements **simultaneous separation and localization**.
Separation Principle

- Additional **latent random variables** need be estimated, to assign each spectrogram point to a single source:

  \[ W_{fl} = m \quad \text{if source } m \text{ emits at } (f, l), \quad \forall 1 \leq f \leq D, 1 \leq l \leq L. \]

- Hence, a matrix \( W = [W_{fl}] \) is estimated for the mixed spectrogram \( S \) (the input).

- From \( W \), binary masks \( \chi_1, \ldots, \chi_m, \ldots, \chi_M \) are evaluated for each source, thus allowing the **mixture spectrogram** \( S \) to be split into \( M \) **source spectrograms** \( S_1, \ldots, S_m, \ldots, S_M \) (the output).

- The inverse short-time Fourier transform (iSTFT) is applied to each spectrogram \( S_m \) to extract the acoustic signal \( a_m(t) \) of source \( m \).
Localization Principle

- Previously we studied a localization method applied to a single source. We estimated $p(x|S; \tilde{\theta})$ given the spectrogram $S = \{s_1, \ldots, s_l, \ldots, s_L, \Lambda\}$ and the model parameters $\tilde{\theta}$.
- In case of $M$ sources, the same localization method can be used, namely evaluation of $p(x_m|S_m; \tilde{\theta})$ with $S_m = \{s_1, \ldots, s_l, \ldots, s_L, \chi_m\}$ and:

$$\tilde{x} = E[x_m|S_m; \tilde{\theta}]$$
Probabilistic Model

- The source assignment variables $W$ are independent random variables $p(W) = \prod_{f,l} p(W_{fl})$. Let:

$$p(W_{fl} = m) = \rho_{fm}, \quad \text{with} \quad \sum_{m=1}^{M} \rho_{fm} = 1, \quad \rho = \{\rho_{fm}\}_{f=1, l=1}^{f=D, l=L}$$

$\rho_{fm}$ is the probability of presence (or proportion) of source $m$ at frequency $f$.

- The problem can now be stated as an expectation-maximization formulation:

$$\tilde{\rho} = \arg\max_{\rho} \mathbb{E}[\log p(S, X, Z, W; \tilde{\theta}, \rho)]$$
Variational EM Algorithm

- **E-step separation:** Evaluate the posterior probability \( p(W|S, X, Z; \tilde{\theta}, \rho^{(i)}) \)
- **E-step localization:** Evaluate the posterior probability \( p(X|S, Z, W; \tilde{\theta}, \rho^{(i)}) \)
- **M-step:**
  \[
  \rho^{(i+1)} = \operatorname{argmax}_\rho \mathbb{E}[ \log p(S, X, Z, W; \tilde{\theta}, \rho, \rho^{(i)}) ]
  \]
Separation and Localization Example

(a) Mixed ILPD spectrogram
(b) direction of source #1  (c) direction of source #2
(d) mask of source #1  (e) mask of source #2
(f) and (g) ground-truth masks of source #1 and source #2
Session Summary

- Sound-source separation amounts to estimate binary masks
- Source separation and localization can be combined in a single model
- A variational EM algorithm was briefly described
- Binaural spectrograms contain very rich information
Week Summary

- Build a spectrogram using ILD and IPD binaural features
- Binaural features encode sound direction
- How to collect data
- Analyzing the binaural features using manifold learning
Week Summary (Continued)

- Non-parametric sound-source localization
- Regression techniques
- Principles of probabilistic localization and separation
- Localizing a single speaker in an image