Audio source separation

based on the sparsity and spatial diversity of the source signals

Simon Leglaive

CentraleSupélec

Today

How to exploit the structure of audio/speech signals to solve the audio source separation problem.

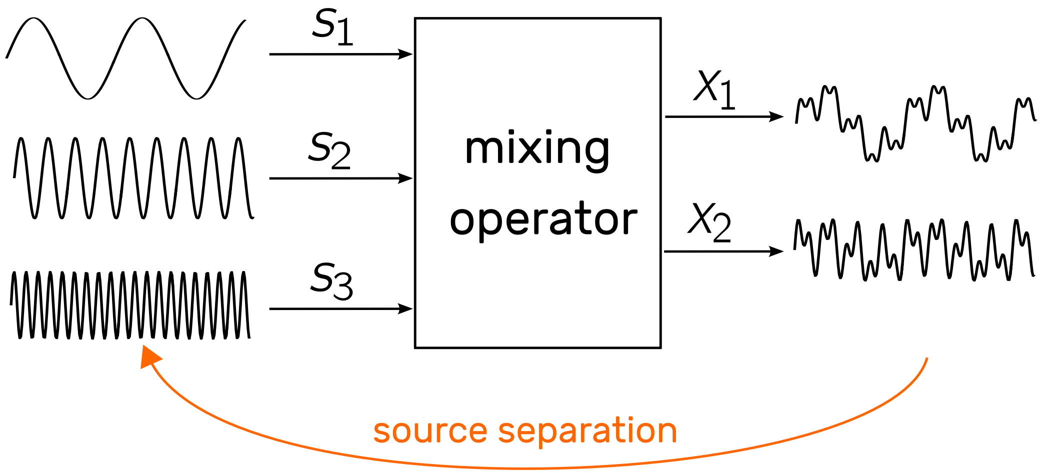
General objectives

The audio source separation problem today is a pretext to discuss general methodological principles involved in many different signal processing problems:

1. The observation model to relate the latent signal(s) of interest to the observations;
2. The latent signal model to make the resolution of the problem tractable;
3. The algorithm to estimate the model parameters and recover the latent signal(s).

It will also be an opportunity to reuse some concepts that we have seen during the last lesson about signal representations. In particular how signal transformations can be useful to define models.

Source separation



The goal is to separate a set of J source signals $s_j(t)$, $j \in \{1, ..., J\}$,
given a set of I mixture signals $x_i(t)$, $i \in \{1, ..., I\}$.

The source separation problem is mainly characterized by

- the type of mixing (instantaneous vs. convolutive);
- the relative number of sources and microphones (under/over-determined problem).

Observation model

Linear instantaneous model

- We aim to model the relationship between the latent signals of interest (the sources) and the observations (the mixtures).

- The simplest mixture model is linear and instantaneous:

$$x_i(t) = \sum_{j=1}^{J} a_{ij}\, s_j(t),$$

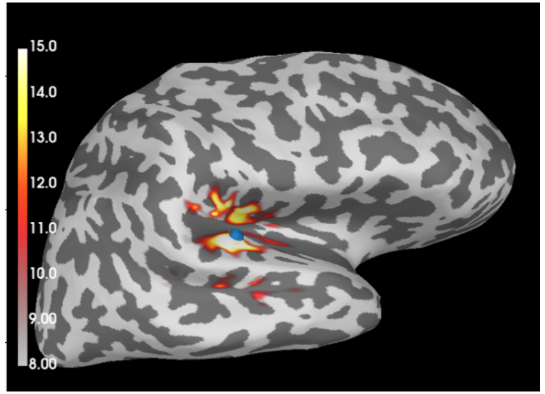
or equivalently in matrix form:

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_I(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1J} \\ a_{21} & a_{22} & \dots & a_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ a_{I1} & a_{I2} & \dots & a_{IJ} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_J(t) \end{bmatrix}.$$

- Magnetoencephalography and electroencephalography (M/EEG) are non-invasive techniques to record brain activity.

  They capture the magnetic and electric signals produced by active neurons from the scalp surface.

- Each M/EEG sensor captures a linear combination of the different brain activities.
- The linear combination is considered instantaneous due to the proximity between the brain and the sensors.



Credits: MNE Python

Credits: Hulton-Deutsch / Corbis Historical / Getty Images

- An electrocardiogram (ECG) is a recording of the heart's electrical activity through repeated cardiac cycles.

- The fetal ECG provides important information about the health of the fetus. Its extraction involves the elimination of the maternal ECG components and other interfering signals from the ECG measurements obtained during pregnancy.

- This can be formulated as a source separation problem, where the mixture is usually assumed linear and instantaneous.

heart.

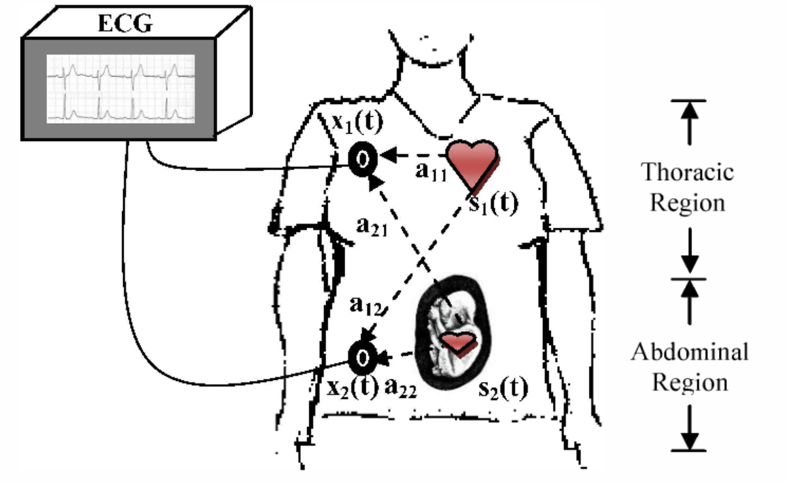
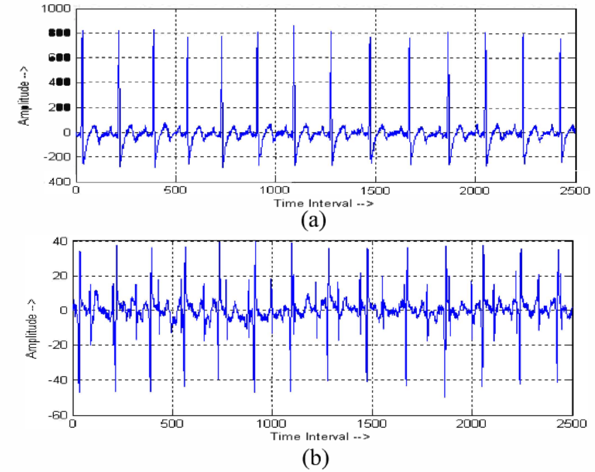


Figure 2: Placement of the electrodes



(a)

(b)

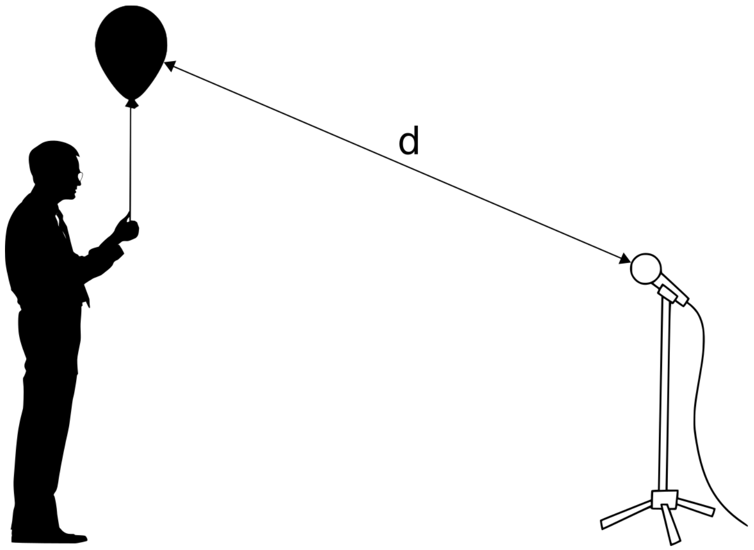
Figure 5: (a) Thoracic Signal (b) Abdominal Signal

Credits: D. Sugumar et al., Joint blind source separation algorithms in the separation of non-invasive maternal and fetal ECG, IECS, 2014.

Anechoic model

In audio, the instantaneous mixing model rarely holds due to the propagation of the sound source in the acoustic medium.

- Let us consider a source (baloon) and a microphone in an open air environment without any obstacle.
- We assume the source and the microphone are not reflecting sound.
- The baloon explodes, it produces a source signal $s(t)$.



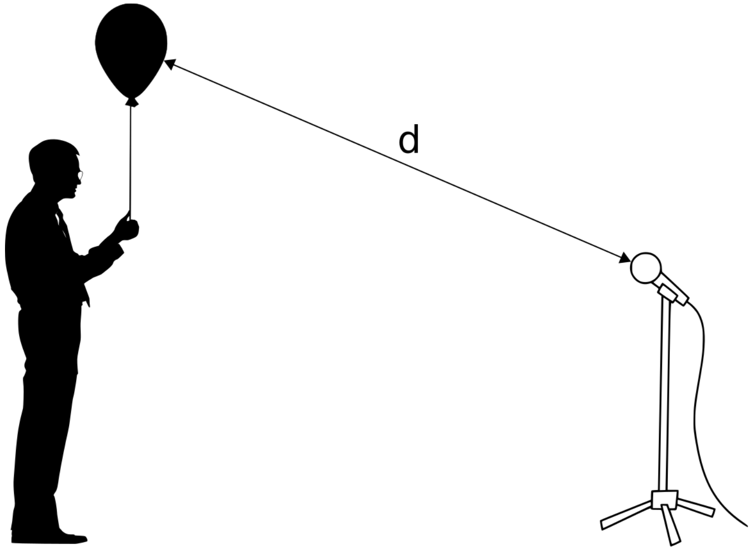
How is the source signal $s(t)$ related to the microphone signal $x(t)$?

The signal acquired by the microphone is given by

$$x(t) = \frac{1}{\sqrt{4\pi d}} s\left(t - \frac{d}{c} f_s\right)$$

where

- d is the source-to-microphone distance (in m);
- $c = 343$ is the speed of sound (in m/s at 20°C);
- d/c is the time of arrival (in s);
- f_s is the sampling rate (in Hz).



At the microphone, the source signal is simply attenuated and delayed.
This is an anechoic recording.

- In a real recording situation, the acoustic environment includes obstacles that affect the sound propagation in many different ways.

- The interaction between the source signal and the acoustic environment is what leads to signal at the microphone.

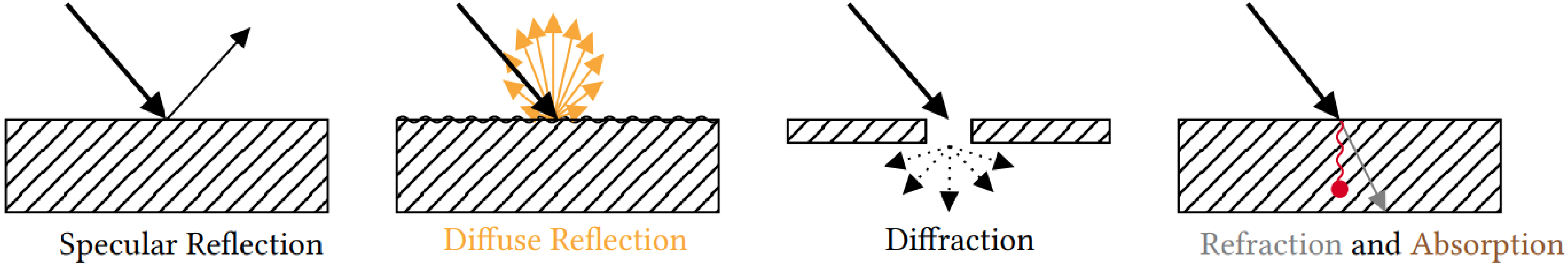


Image credits: Diego Di Carlo's Ph.D. Thesis "Echo-aware signal processing for audio scene analysis", 2020.

Convolutive model

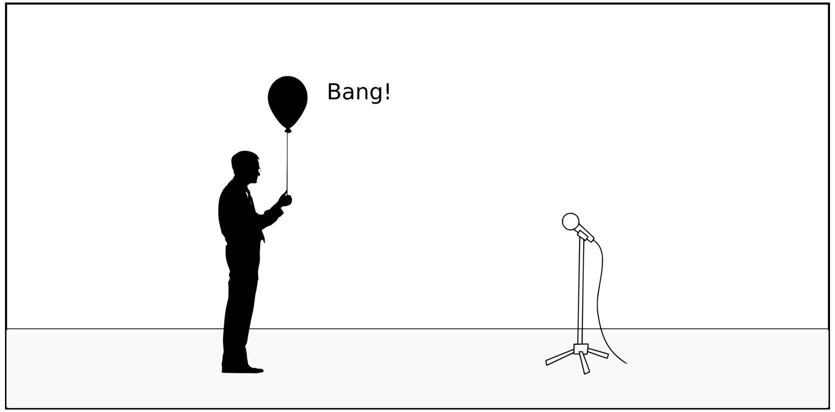
- This interaction is accurately represented by a convolution:

$$x(t) = [h \star s](t) = \sum_{\tau=0}^{L_h-1} h(\tau)s(t-\tau),$$

where $h(t)$ is called the room impulse response (RIR) and characterizes the acoustic path between the source and microphone locations.

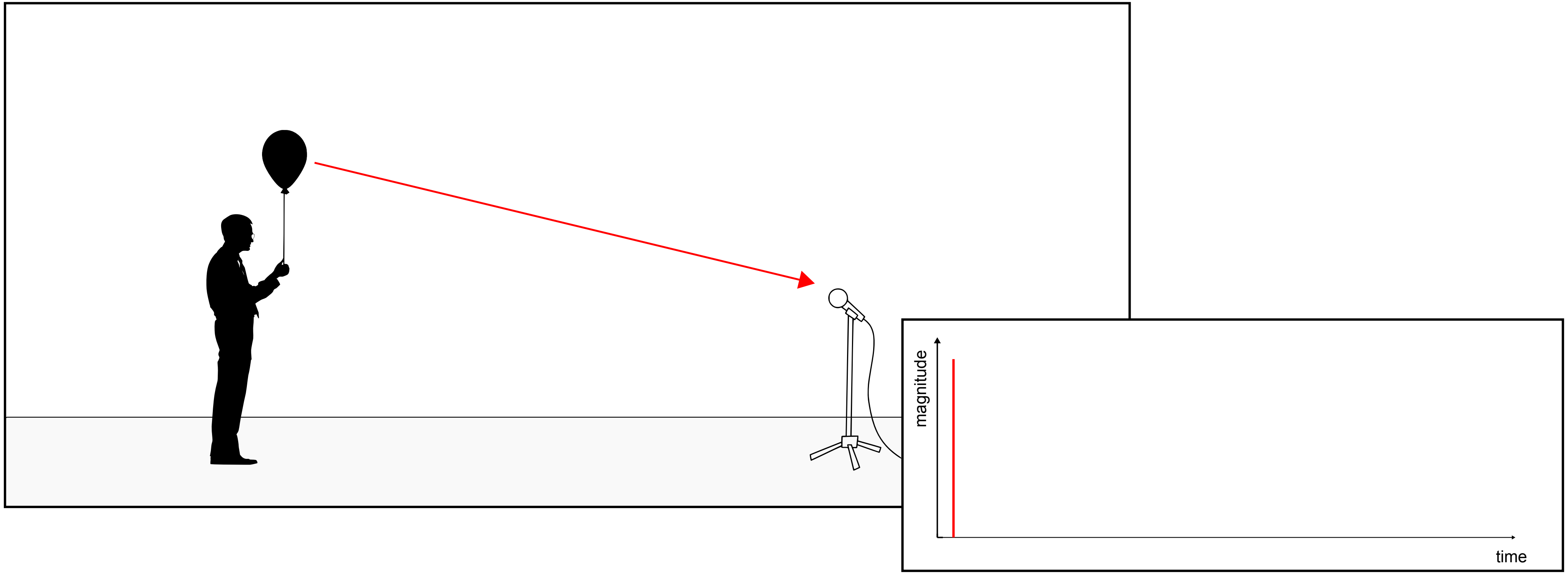
- It is called the room *impulse* response because it is the response of the room when the source is an impulse (dirac delta function):

$$x(t) = [h \star \delta](t) = h(t).$$



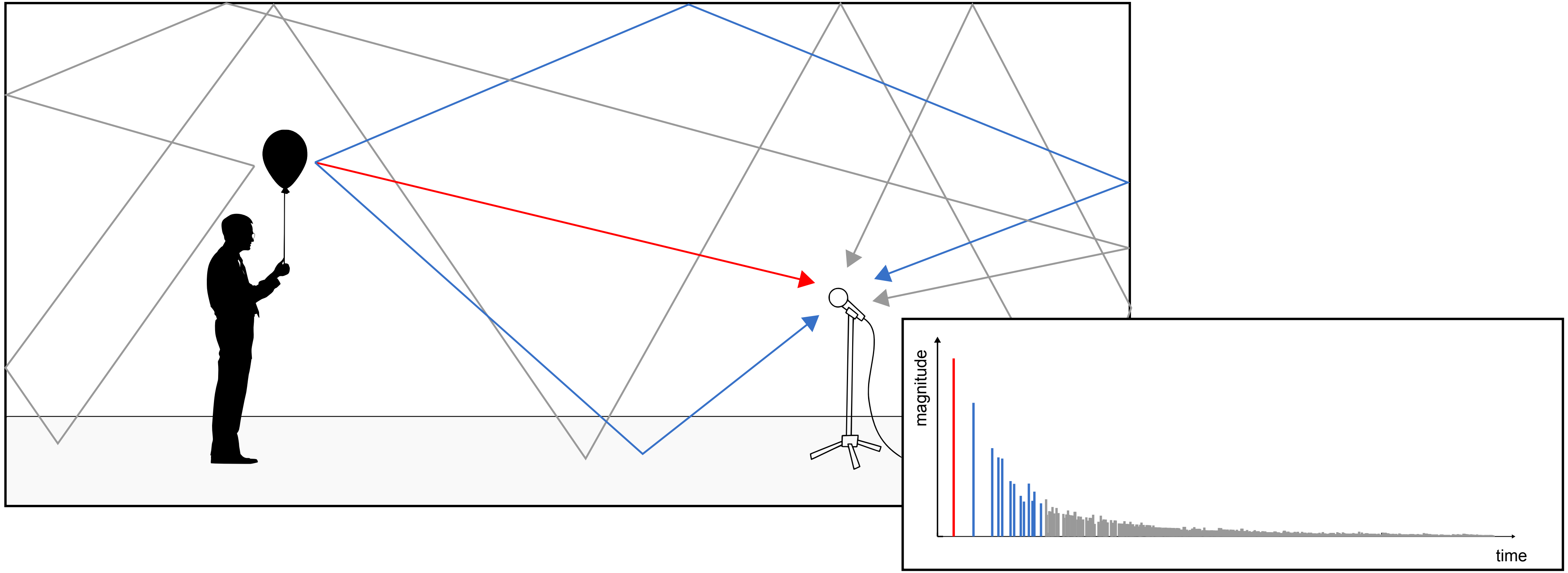
The previous anechoic model corresponds to the case where the RIR only characterizes the direct path between the source and the microphone:

$$h(t) = \frac{1}{\sqrt{4\pi d}}\delta\left(t - \frac{d}{c}f_s\right), \qquad \delta(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases}.$$



But in a real room, many reflections of the sound source arrive at the microphone.

This is called reverberation.



Summary

- To sum up, we have seen different mixture models:
  - Linear instantaneous mixture model: $$x_i(t) = \sum_{j=1}^{J} a_{ij}\, s_j(t).$$
  - Anechoic mixture model: $$x_i(t) = \sum_{j=1}^{J} a_{ij}\, s_j(t - \tau_{ij}).$$
  - Convolutive mixture model: $$x_i(t) = \sum_{j=1}^{J} [a_{ij} \star s_j](t).$$
- The more "expressive" the mixture model, the more complex in terms of the number of unknown mixing parameters.

  We are not directly interested in the mixing parameters, but we will have to estimate them to solve the source separation problem.

  More unkowns means a more difficult problem to solve. There is in general a trade-off between modeling accuracy and estimation tractability.

Latent source signal model

Blind under-determined source separation

For simplicty, let us consider the linear instantaneous model:

$$\begin{bmatrix} x_1(t) \\ \vdots \\ x_I(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1J} \\ a_{21} & a_{22} & \dots & a_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ a_{I1} & a_{I2} & \dots & a_{IJ} \end{bmatrix} \begin{bmatrix} s_1(t) \\ s_2(t) \\ \vdots \\ s_J(t) \end{bmatrix}.$$

Moreover, we consider an under-determined scenario where the number of microphones I is lower than the number of sources J, i.e. we have more unkowns than equations.

This is an ill-posed problem that admits an infinite number of solution.

Regularization with a signal model

- We need to introduce additional information about the latent signals of interest to compensate for the lack of observations.

- This additional information will be provided through a source signal model, which can take different forms:

  - simplifying assumptions (e.g., the sources have disjoint time supports);

  - deterministic model (e.g., $s_j(t)$ is the sum of a few sinusoids with exponentially decaying amplitudes);

  - probabilistic model (e.g., $s_j(t)$ is a locally stationary Gaussian process).

Signal models are usually expressed mathematically.
Using mathematical representations of signals allows us to derive algorithms to solve real-world problems (e.g., source separation).

Signal modeling in a transformed domain

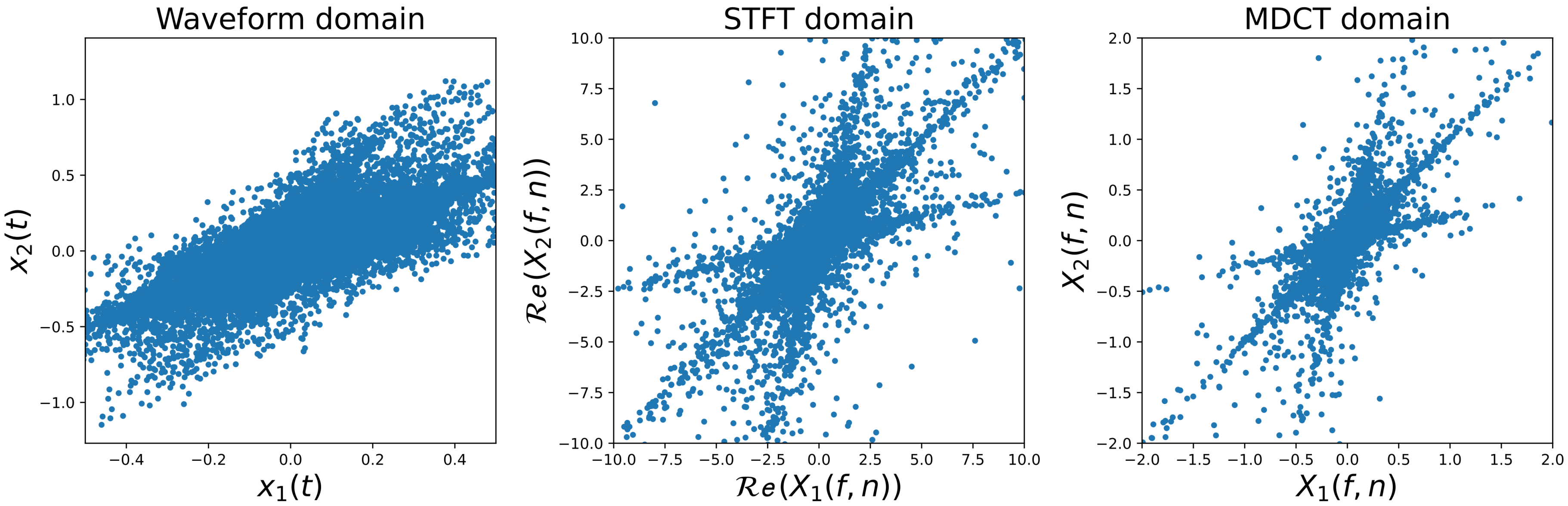
- Sometimes, it is easier to define a signal model in a transformed domain.
- We are processing non stationary audio signals, so we consider a linear time-frequency transform (see the previous lesson on signal representations), such as the MDCT or the STFT.
- The observation model in the transformed domain is simply given by

$$\begin{bmatrix} X_1(f,n) \\ \vdots \\ X_I(f,n) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1J} \\ a_{21} & a_{22} & \dots & a_{2J} \\ \vdots & \vdots & \vdots & \vdots \\ a_{I1} & a_{I2} & \dots & a_{IJ} \end{bmatrix} \begin{bmatrix} S_1(f,n) \\ S_2(f,n) \\ \vdots \\ S_J(f,n) \end{bmatrix},$$

where $\cdot(f,n)$ denotes a signal coefficient in the STFT or MDCT domain, at the time-frequency point $(f,n) \in \{0, ..., F-1\} \times \{0, ..., N-1\}$.

Let us consider a mixture of $J = 3$ speech sources over $I = 2$ microphones, represented in the waveform, MDCT or STFT domain by:

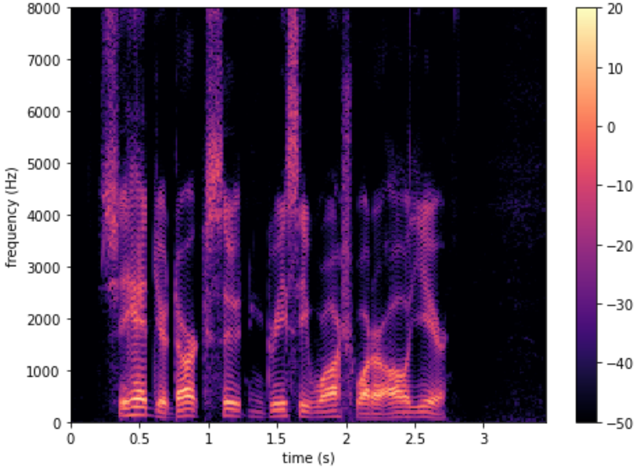
$$\begin{bmatrix} x_1(\cdot) \\ x_2(\cdot) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} s_1(\cdot) \\ s_2(\cdot) \\ s_3(\cdot) \end{bmatrix} = s_1(\cdot) \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} + s_2(\cdot) \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} + s_3(\cdot) \begin{bmatrix} a_{13} \\ a_{23} \end{bmatrix}.$$



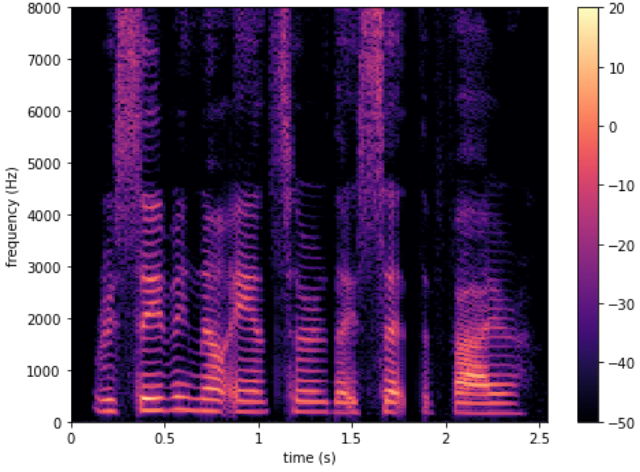
The figure shows the mixture coefficients in the different domains, what do you observe and why?

- We see that some structure emerges in the MDCT/STFT representation of the mixture signals.
- This structure actually originates from the source signals, which tend to be sparse in the time-frequency domain.
- Sparsity is a central notion in signal processing, and we can define signal models that encode this characteristic of natural signals and images.

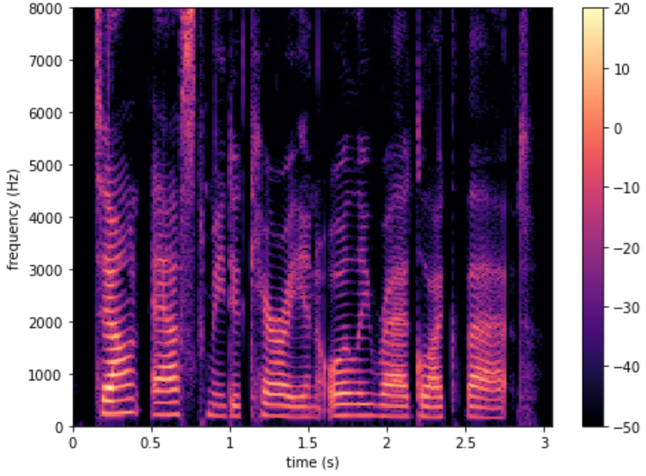
Source 1



Source 2



Source 3



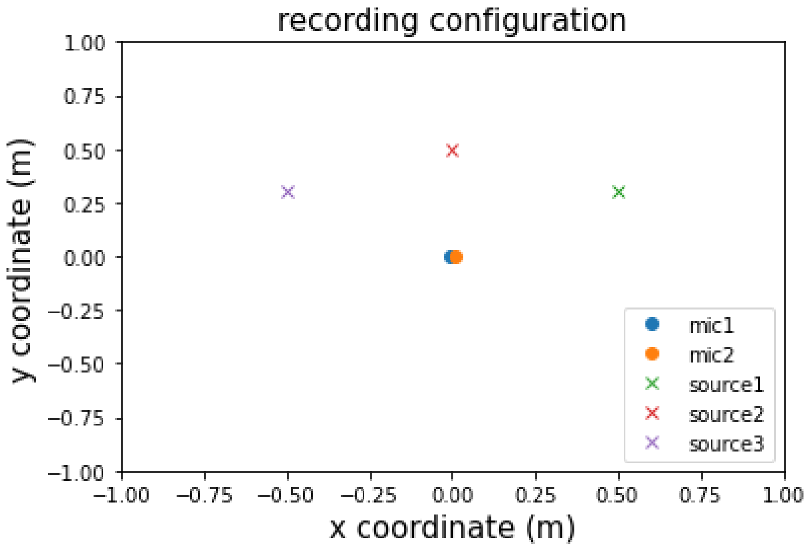
The Degenerate Unmixing Estimation Technique (DUET)

We are now going see that sparsity is used in the DUET algorithm to solve the audio source separation problem for anechoic stereophonic mixtures.

S. Rickard, "The DUET Blind Source Separation Algorithm", 2007.

Unmixing or source separation

We want to estimate individual speech source signals from a stereophonic mixture.



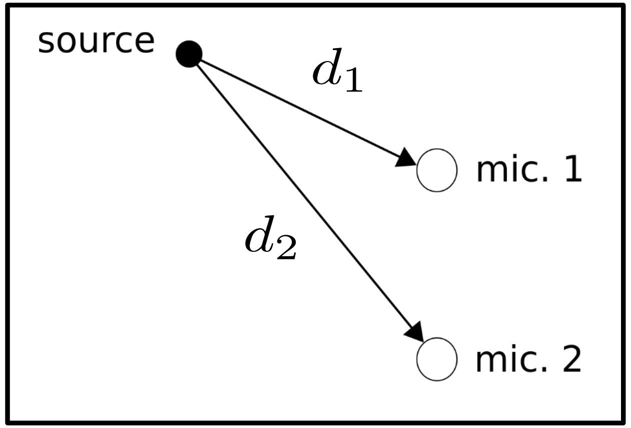
We have 3 source signals so the problem is under-determined or degenerate.

Anechoic stereo mixture model in the time domain

Each microphone signal is the sum of delayed and attenuated source signals:

$$x_i(t) = \sum_{j=1}^{J} \frac{1}{\sqrt{4\pi} d_{ij}} s_j \left(t - \frac{d_{ij}}{c} f_s \right),$$

where d_{ij} is the distance between the j-th source and the i-th microphone



Without loss of generality, we absorb the attenuation and delay parameters at the first microphone into the definition of the source signal, i.e., we define the "new" source signal by:

$$\tilde{s}_j(t) = \frac{1}{\sqrt{4\pi d_{1j}}} s_j\left(t - \frac{d_{1j}}{c} f_s\right)$$

such that the mixture model becomes

$$x_1(t) = \sum_{j=1}^{J} \tilde{s}_j(t), \qquad x_2(t) \quad = \sum_{j=1}^{J} a_j \tilde{s}_j(t - \delta_j),$$

where

- $a_j = d_{1j}/d_{2j}$ is the relative attenuation factor of the j-th source, also called the inter-microphone level ratio;
- $\delta_j = \dfrac{d_{2j} - d_{1j}}{c} f_s$ is the time difference of arrival (TDoA) of the j-th source, also called the inter-microphone time difference.

These parameters convey information about the spatial location of the source.

In matrix form, we have:

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ a_{21} & a_{22} & \dots & a_{2J} \end{bmatrix} \begin{bmatrix} s_1(t-\delta_1) \\ s_2(t-\delta_2) \\ \vdots \\ s_J(t-\delta_J) \end{bmatrix},$$

where we removed the tilde to simplify the notations.

In the following, we will refer to $\{(a_j, \delta_j)\}_{j=1}^{J}$ as the mixing parameters.

Anechoic stereo mixture model in the STFT domain

Assuming that the TDoAs are small relative to the STFT analysis window length L, we have:

$$s_j(t-\delta_j) \overset{\text{STFT}}{\longleftrightarrow} \exp\left(-\imath 2\pi \frac{f\delta_j}{L}\right) S_j(f,n).$$

The mixture model thus rewrites in STFT domain as follows:

$$\begin{bmatrix} X_1(f,n) \\ X_2(f,n) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 \exp\left(-\imath 2\pi \frac{f\delta_1}{L}\right) & \dots & a_J \exp\left(-\imath 2\pi \frac{f\delta_J}{L}\right) \end{bmatrix} \begin{bmatrix} S_1(f,n) \\ \vdots \\ S_J(f,n) \end{bmatrix}.$$

DUET principle

It is possible to blindly separate an arbitrary number of sources from **two anechoic mixtures** provided that

- the **time–frequency representations of the sources do not overlap** (assumption 1),
- the sources have **different spatial locations** (assumption 2).

W-disjoint orthogonality (assumption 1)

- Source signals are sparse and have disjoint time-frequency supports. In other words, at most one source is active at each time-frequency point (f,n).

- The W-disjoint orthogonality hypothesis can be formalized by:

$$S_j(f,n)S_k(f,n) = 0, \qquad \forall (f,n), \qquad \forall j \neq k.$$

- The mixture model simplifies as follows:

$$\begin{bmatrix} X_1(f,n) \\ X_2(f,n) \end{bmatrix} = \begin{bmatrix} 1 \\ a_{\mathcal{I}(f,n)} \exp\left(-\imath 2\pi \frac{f \delta_{\mathcal{I}(f,n)}}{L} \right) \end{bmatrix} S_{\mathcal{I}(f,n)}(f,n).$$

where $\mathcal{I}(f,n) \in \{1, 2, ..., J\}$ indicates which source is active at time-frequency point (f,n).

- It is the mathematical idealization of a milder assumption considering that every time–frequency point in the mixture is dominated by the contribution of at most one source.

Unmixing with binary masking

W-disjoint orthogonality is crucial to DUET because it allows for separating the mixture into its component sources using binary masks:

$$\hat{S}_j(f,n) = M_j(f,n)X_1(f,n),$$

where the mask is defined by:

$$M_j(f,n) = \begin{cases} 1 & \text{if } \mathcal{I}(f,n) = j \\ 0 & \text{otherwise} \end{cases}.$$

The source separation problem now becomes that of estimating which source is active at each time-frequency point.

This is where the second assumption of DUET comes into play.

DUET algorithm

Let us recall the mixture model under the W-disjoint orthogonality assumption:

$$\begin{bmatrix} X_1(f,n) \\ X_2(f,n) \end{bmatrix} = \begin{bmatrix} 1 \\ a_{\mathcal{I}(f,n)} \exp\left(-\imath 2\pi \frac{f \delta_{\mathcal{I}(f,n)}}{L} \right) \end{bmatrix} S_{\mathcal{I}(f,n)}(f,n).$$

The main observation that DUET leverages is that **the ratio of the mixtures in the STFT domain does not depend on the source signal but only on the mixing parameters associated with the active source**:

$$\frac{X_2(f,n)}{X_1(f,n)} = a_j \exp\left(-\imath 2\pi \frac{f \delta_j}{L} \right), \qquad \forall (f,n) \in \Omega_j = \{(f,n), \mathcal{I}(f,n) = j\}.$$

- Let us define the local attenuations and delays by:

$$\hat{a}(f,n) = \left| \frac{X_2(f,n)}{X_1(f,n)} \right|,$$

$$\hat{\delta}(f,n) = -\frac{1}{2\pi f/L} \arg\left(\frac{X_2(f,n)}{X_1(f,n)} \right), \qquad f > 0.$$

- Using the key observation in the previous slide, we have:

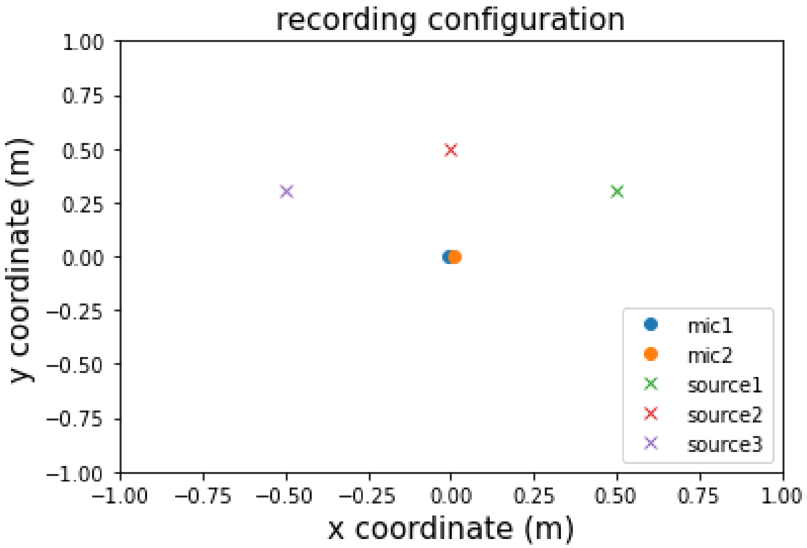
$$\left(\hat{a}(f,n), \hat{\delta}(f,n)\right) = \left(a_j, \delta_j\right), \qquad \forall (f,n) \in \Omega_j = \{(f,n), \mathcal{I}(f,n) = j\}.$$

Spatial diversity (assumption 2)

We assume that the **sources have different spatial locations**, that is

$$(a_j \neq a_k) \text{ or } (\delta_j \neq \delta_k), \qquad \forall j \neq k.$$

We recall that a_j and δ_j encode the position of the j-th source relative to the microphones.

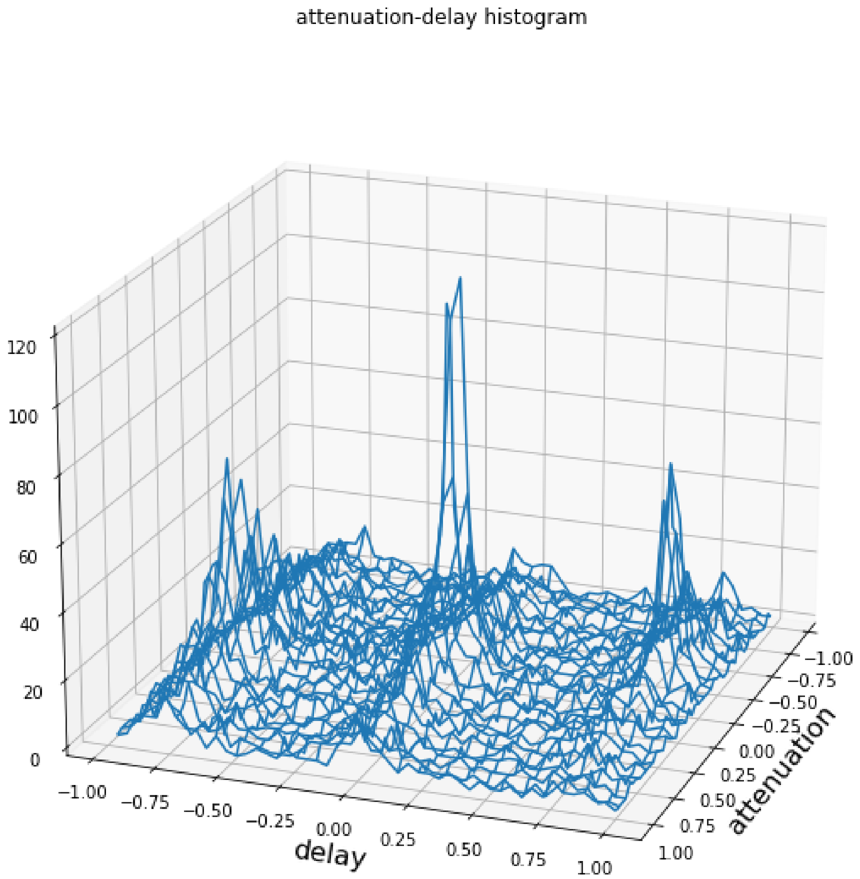


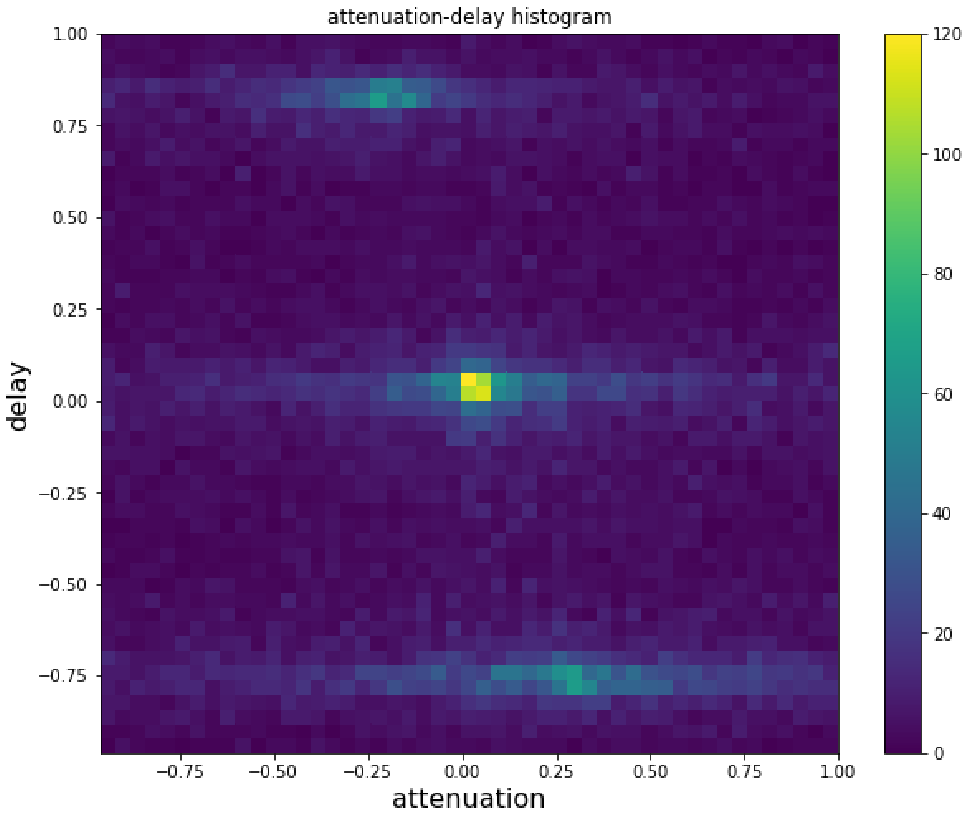
2D histogram of local attenuations and delays

The local attenuations and delays that are computed from the mixture signals can thus only take values among the actual mixing parameters that are assumed to be all different:

$$\big(\hat{a}(f,n), \hat{\delta}(f,n)\big) \in \{(a_j, \delta_j)\}_{j=1}^{J}, \qquad \forall (f,n).$$

What should we obtain if we build a 2D histogram of the local attenuations and delays $\big(\hat{a}(f,n), \hat{\delta}(f,n)\big)$?





The observations do not perfectly match with the model, but we can still identify three clusters.

- From the 2D histogram, we can estimate the mixing parameters $\{(\hat{a}_j, \hat{\delta}_j)\}_{j=1}^J$ by peak picking.
- We recall that in principle, for all time-frequency points (f, n),

$$\big(\hat{a}(f,n), \hat{\delta}(f,n)\big) \in \{(\hat{a}_j, \hat{\delta}_j)\}_{j=1}^J.$$

- We can thus build the time-frequency masks for source separation as follows:

$$M_j(f,n) = \begin{cases} 1 & \text{if } \big(\hat{a}(f,n), \hat{\delta}(f,n)\big) = (\hat{a}_j, \hat{\delta}_j) \\ 0 & \text{otherwise} \end{cases}.$$

- In practice, because not all the assumptions are strictly satisfied, the local attenuations and delays will not be precisely equal to the estimated mixing parameters, but they will cluster around them. We will need a metric to measure the proximity.

Summary of DUET

1. Construct the STFT representations $X_1(f,n)$ and $X_2(f,n)$ of both mixtures.

2. Take the ratio of the two mixtures and extract local attenuations and delays

$$\left\{\left(\hat{a}(f,n), \hat{\delta}(f,n)\right)\right\}_{(f,n)}.$$

1. Compute a 2D histogram and estimate the mixing parameters $\left\{(\hat{a}_j, \hat{\delta}_j)\right\}_{j=1}^{J}$ by peak picking.

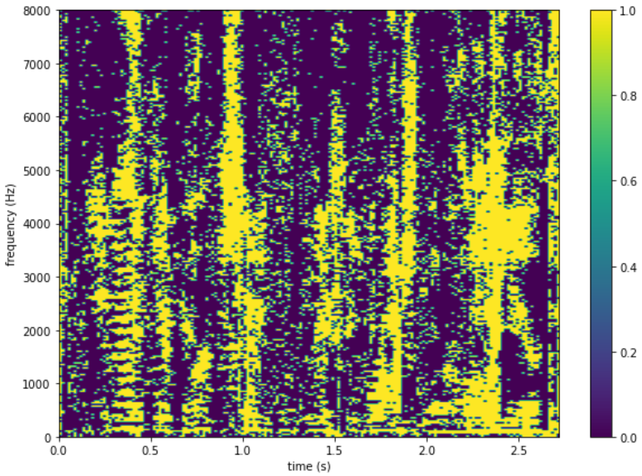
2. Build the binary masks

$$M_j(f,n) = \begin{cases} 1 & \text{if } \left(\hat{a}(f,n), \hat{\delta}(f,n)\right) \approx (\hat{a}_j, \hat{\delta}_j) \\ 0 & \text{otherwise} \end{cases}.$$

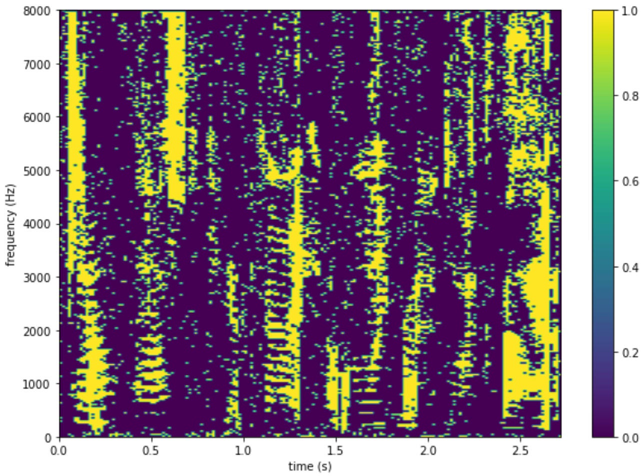
3. Estimate the sources by $\hat{S}_j(f,n) = M_j(f,n) X_1(f,n)$.

4. Compute the inverse STFT to get the time-domain source signals.

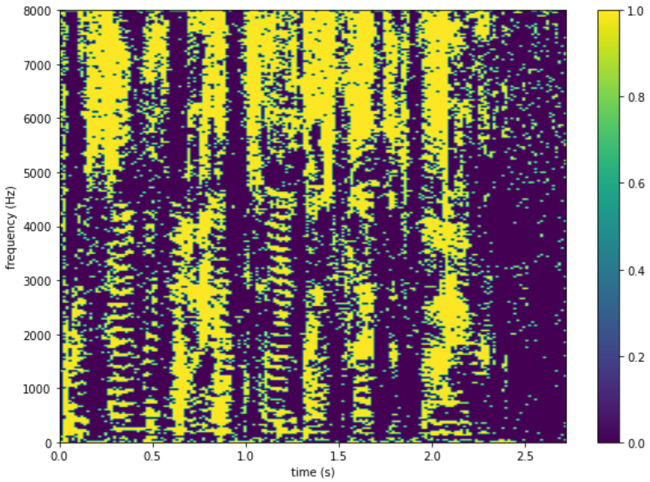
Mask 1



Mask 2



Mask 3


frequency (Hz)
time (s)