Lecture 7: Multi-channel Acoustic Signal Processing

Advanced Signal Processing / Music and Speech Signal Processing

Kouei Yamaoka

Graduate School of Information Science and Technology, The University of Tokyo
Assistant Professor, Saruwatari & Saito Laboratory

Lecture Schedule

04/07: Lecture 1 Statistical speech & acoustic signal processing
04/14: Lecture 2 Nonnegative matrix factorization
04/21: Lecture 3 Blind source separation 1
04/28: Lecture 4 Blind source separation 2
05/12: Lecture 5 Enhancement & higher-order statistical
05/19: Lecture 6 [Report-1]
05/26: Lecture 7 Multi-channel acoustic signal processing
06/09: Lecture 8 Music information processing
06/16: Lecture 9 Environmental sound generation & recognition
06/23: Lecture 10 Speech synthesis & voice conversion
06/30: Lecture 11 [Report-2]

Report for the Second-Half Lectures

- **Assignment**
  - Select **three** international conference papers related to Lectures 7–10
  - Summarize each paper in about **three slides**
  - Papers: papers published in top conferences in speech/acoustic signal processing from the last five years, such as ICASSP or INTERSPEECH
- **Submission format**
  - Submit your report as a **PDF**. The cover page must include: course title, department, year, student ID, and name
- **Deadline**
  - **July 7**. The submission page will be available on UTOL

Self-Introduction

- **Name:** YAMAOKA Kouei (Email: kouei_yamaoka@ipc.i.u-tokyo.ac.jp)
- **Background**
  - 2019: M.Eng., Graduate School of Systems and Information Engineering, University of Tsukuba
  - 2024: Ph.D. in Information Science, Graduate School of System Design, Tokyo Metropolitan University
  - 2024–present: Assistant Professor, Saruwatari & Saito Laboratory
- **Research interest**
  - Multi-channel acoustic signal processing, distributed microphone array, beamforming, time delay estimation

Outline of This Lecture

- **Overview of multi-channel acoustic signal processing**
  - Tasks in multi-channel acoustic signal processing
  - Physics in multi-channel acoustic signal processing
- **Time delay estimation**
  - Overview of time delay estimation
  - Observation models for time delay estimation
  - Optimization: Cross-correlation and maximum likelihood estimation
  - Theoretical performance analysis
  - Numerical evaluation
- **Summary**

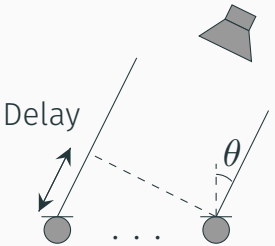
Overview of Multi-channel Acoustic Signal Processing

What Is Signal Processing?

- “A branch of electrical engineering which pulls meaning from the broad sources of data all around us”
  - Quoted from the IEEE Signal Processing Society website
  - What is Signal Processing? (`https://signalprocessingsociety.org/our-story/signal-processing-101`)
  - YouTube: `https://www.youtube.com/watch?v=R90ciUoxcJU`
- Signal processing 1. models signals and 2. analyzes them for a desired purpose
  - Targets: acoustics, images, electricity, biological signals, etc.
  - Applications: computers, TVs, telephones, mobile devices, and more

Why Multi-channel?

- Waves propagate through space
  - Observations depend on sensor position
- Multiple sensors allow us to estimate spatial properties of waves
  - Where does the wave come from?
  - Which direction should be enhanced?
  - How many sources are present?



▶ Sensor arrays are a fundamental tool for exploiting spatial information

Mutli-channel Acoustic Signal Processing

- **Microphone array signal processing (MASP)**
  - Signal processing for signals captured by multiple microphones
  - Uses not only temporal and spectral information
  - But also spatial information in the observations



RWCP-SSD [S. Nakamura+, 2002]



Meeting Owl [meetingowl.jp]



Sony ICD-SX813

Applications of MASP

Voice activity detection

Speaker tracking

Source localization



Signal enhancement

Noise reduction

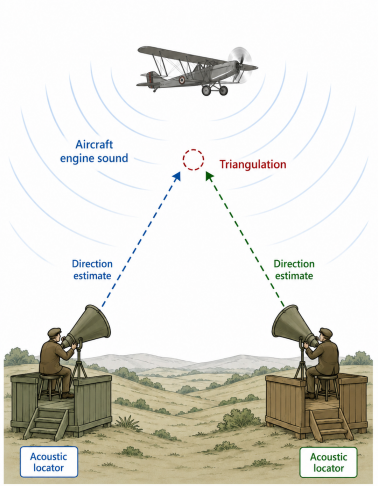
Echo cancellation

Meeting Owl [meetingowl.jp]
Equipped with an 8-channel microphone array

- **Goal:**
  computationally emulate, enhance, and extend human auditory capabilities

Early Acoustic Sensing

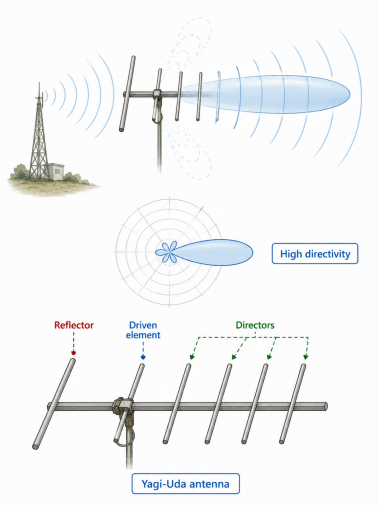
- **Aircraft detection by sound**
  - Used from the 1910s
  - Acoustic locators collect engine sounds
  - Listeners estimate arrival directions
  - Multiple direction estimates enable triangulation
- **Relation to sensor arrays**
  - Multiple listening points provided spatial information
  - Modern arrays replace human listeners with microphones and algorithms



Generated by ChatGPT

Antenna Arrays and Directivity

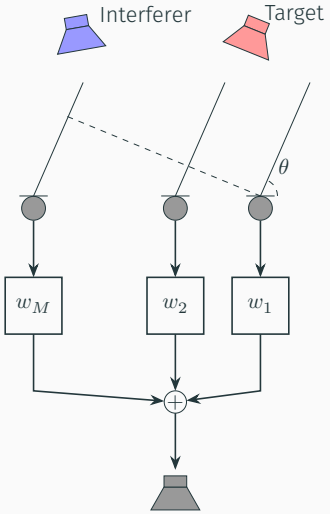
- **Yagi–Uda antenna** [H. Yagi, 1926]
  - A classical example of an antenna array
  - Japan's first *IEEE Milestone*
  - High sensitivity is achieved in a desired direction, and array geometry determines the spatial directivity
- **Relation to sensor arrays**
  - Arrays exploit wave propagation
  - Multiple elements provide spatial selectivity
  - Array processing controls "where to listen"



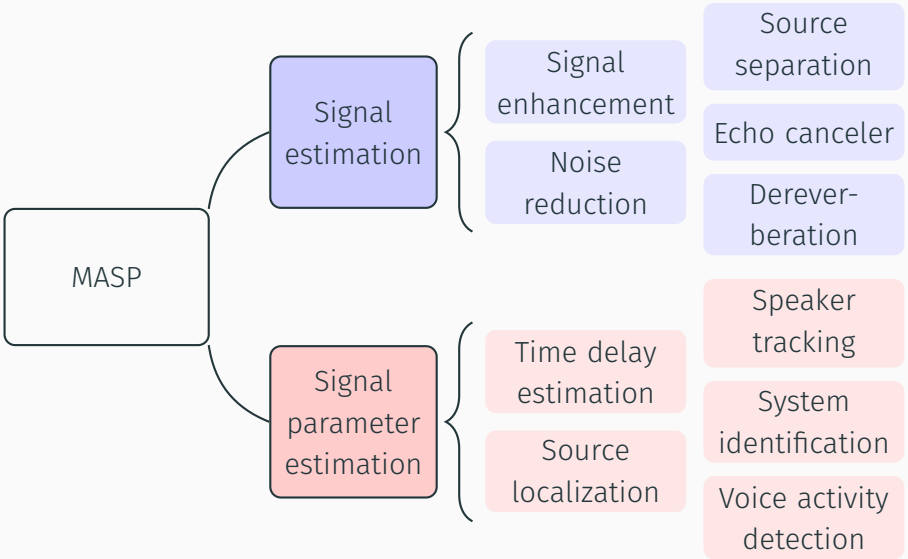
Generated by ChatGPT

From Hardware Directivity to Signal Processing

- **Directivity goes beyond hardware design**
  - It can be controlled by signal processing
- **Basic example: delay-and-sum beamforming**
  - Compensate delays for a target direction
  - Add the aligned signals
  - Advanced methods design spatial filters adaptively via optimization
- **Sensor array processing can be viewed as spatial filtering**

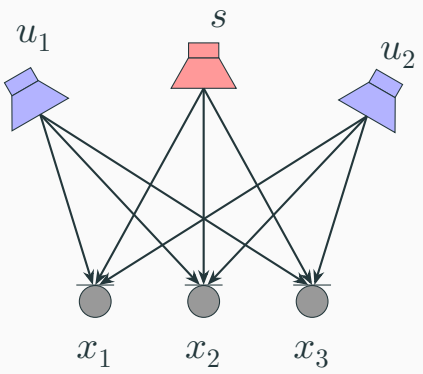


Task Categories in MASP



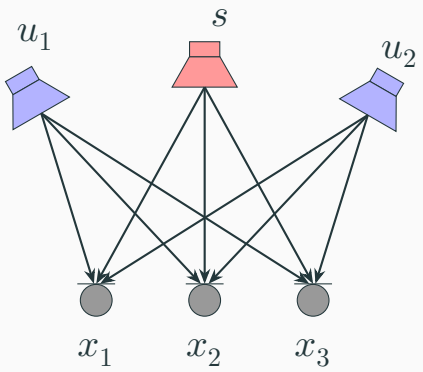
Example of Signal Estimation

- **Signal enhancement**
  - Suppresses noise (u_1, u_2) and enhances the target signal (s)
  - If target is speech: *speech enhancement*
  - Extract multiple sources separately: *source separation*
- **Applications**
  - Robot audition
  - Preprocessing for speech recognition and voice conversion
  - Hearing aids and other assistive devices

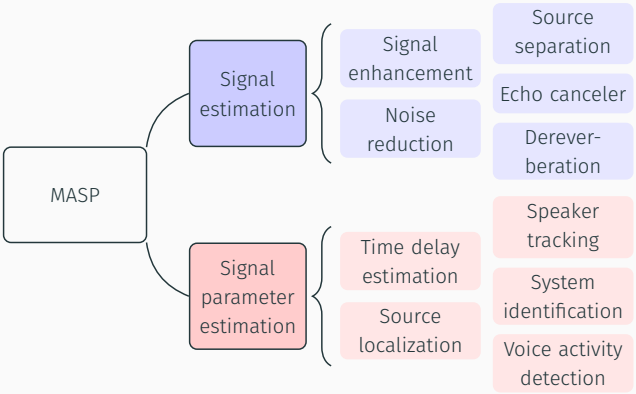


Example of Signal Parameter Estimation

- **Direction of arrival (DOA) estimation**
  - Estimates the direction from which a signal arrives
  - Estimate 3D position: *localization*
  - Estimate source motion over time: *source tracking*
- **Applications**
  - Robot audition
  - Acoustic surveillance
  - Autonomous driving



Topic of this lecture

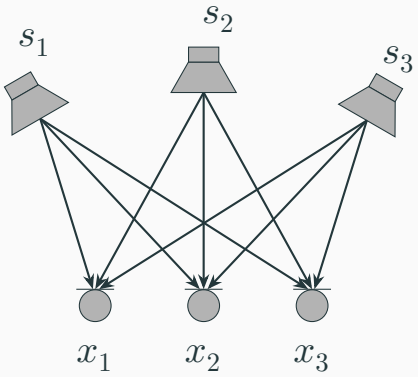


- **Topic of this lecture:** time delay estimation
  - a fundamental tool for parameter estimation

Physical Models in Array Signal Processing

Preliminaries

- **Assumed setup**
  - Source signal: $\breve{s}_q(\tilde{t})$
    (source index $q = 1, \ldots, Q$)
  - Observed signal: $x_m(\tilde{t})$
    (microphone index $m = 1, \ldots, M$)
  - $\tilde{t}$: continuous time variable
- **Assumptions**
  - Microphones are omnidirectional
  - Microphones and speakers behave linearly
  - Microphones and speakers do not move,
    i.e., *linear time-invariant (LTI) system*



Sound Waves and Microphones

- **Sound is a compression wave propagating through air**
  - Air-pressure fluctuations (movement, sequential compression and rarefaction of particles) are the physical quantity of interest
  - Audible sound pressure is roughly $20\,\mu\text{Pa}$ to $20\,\text{Pa}$
- **Microphones transduce sound into electrical signals**
  - This conversion enables us to analyze sounds as signals

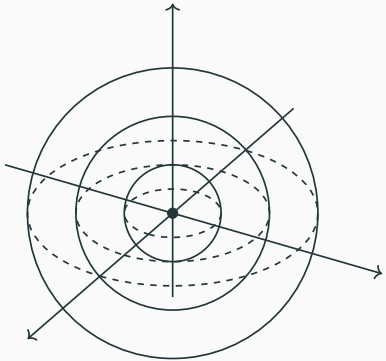
▶ **To model microphone signals, we need a model of sound propagation**

Spherical Wave Model [S. Gannot+, 2018]

- **Sound propagation in a *free field***
  - A point source radiates a Spherical wave
  - The signal is delayed and attenuated according to distance:

$$s_{m,q}(\tilde{t}) = \frac{1}{\sqrt{4\pi} d_{m,q}} \breve{s}_q \left(\tilde{t} - \frac{d_{m,q}}{c} \right)$$

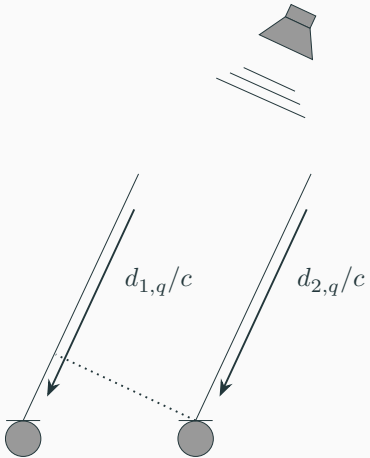
  - Cf. inverse-square law
  - $d_{m,q}$: source–microphone distance, c: speed of sound



Plane Wave Model

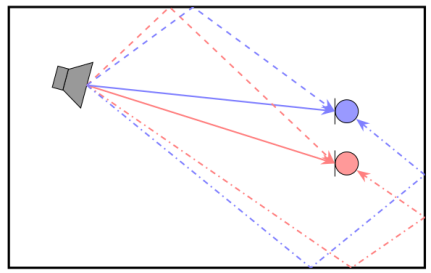
- **Far-field approximation**
  - A distant point source is approximated by a plane wave
  - Wavefronts are almost parallel around the microphone array
  - Attenuation α is approximately identical across microphones
  - Only relative time delays are essential:

$$s_{m,q}(\tilde{t}) = \alpha \breve{s}_q \left(\tilde{t} - \frac{d_{m,q}}{c} \right)$$



Reverberation in Real Environments

- **Sound propagation in a room**
  - Sound reaches microphones through multiple paths
  - The direct sound arrives first
  - Reflected sounds arrive later after bouncing off walls and objects



- *Reverberation*
  - A superposition of many delayed and attenuated sound components
  - Observation depends on the room, source/microphone positions
  - This propagation effect is modeled by an acoustic impulse response

Expression Using Impulse Responses (1/2)

- Sound propagation is written as convolution:

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} h_{1t} * s_t \\ h_{2t} * s_t \end{pmatrix} + \begin{pmatrix} \tilde{u}_{1t} \\ \tilde{u}_{2t} \end{pmatrix}$$

- Observed microphone signals
  - x_{1t}, x_{2t}: observed signals
  - $\tilde{u}_{1t}, \tilde{u}_{2t}$: noise and other undesired components
- Acoustic impulse responses
  - h_{1t}, h_{2t} describe propagation from the source to each microphone

Expression Using Impulse Responses (2/2)

- Sound propagation is written as convolution:

$$\underbrace{\boldsymbol{x}(t)}_{\text{multichannel observation}} = \sum_{q=1}^{Q} \underbrace{\boldsymbol{s}_q(t)}_{\text{source image}}, \qquad \boldsymbol{s}_q(t) = \sum_{\tau} \boldsymbol{h}_q(\tau) \underbrace{\check{s}_q(t-\tau)}_{\text{source signal}}$$

  - $\boldsymbol{x}(t) = [x_1(t) \ \cdots \ x_M(t)]^{\mathsf{T}}$
  - $\boldsymbol{s}_q(t) = \left[s_{1,q}(t) \ \cdots \ s_{M,q}(t)\right]^{\mathsf{T}}$
  - t: discrete-time index

- Acoustic impulse responses
  - $\boldsymbol{h}_q(t) = \left[h_{1,q}(t) \ \cdots \ h_{M,q}(t)\right]^{\mathsf{T}}$
  - Describe direct-path propagation and reverberation

Expression in the Frequency Domain

- In the STFT domain:

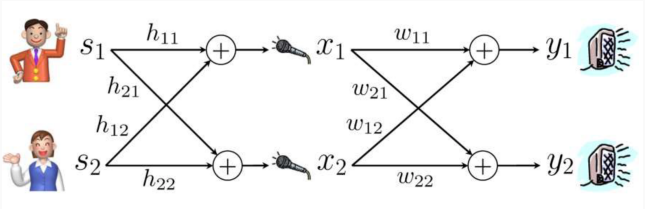
$$\boldsymbol{x}(k,n) = \sum_{q=1}^{Q} \boldsymbol{s}_q(k,n), \qquad \boldsymbol{s}_q(k,n) = \boldsymbol{h}_q(k)\check{s}_q(k,n)$$

  - STFT: Short-time Fourier transform
  - k, n: frequency bin and time-frame indices
  - $\boldsymbol{s}_q(k,n)$ and $\check{s}_q(k,n)$: STFT representations of $\boldsymbol{s}_q(t)$ and $\check{s}_q(t)$
  - Assumption: the frame is sufficiently long (*narrowband approximation*)

- Acoustic transfer functions (*ATFs*)
  - $\boldsymbol{h}_q(k) = \left[h_{1,q}(k) \ \cdots \ h_{M,q}(k)\right]^{\mathsf{T}}$
  - Provides a simple and effective model of multichannel observation

Connection to Source Separation

- **Source separation uses the same ATF-based model**
  - Each source is mixed through propagation paths to microphones



Quoted from the lecture materials for Lectures 3 and 4

$$\begin{pmatrix} x_1(k,n) \\ \vdots \\ x_M(k,n) \end{pmatrix} = \begin{pmatrix} h_{1,1}(k) & \cdots & h_{1,Q}(k) \\ \vdots & \ddots & \vdots \\ h_{M,1}(k) & \cdots & h_{M,Q}(k) \end{pmatrix} \begin{pmatrix} s_1(k,n) \\ \vdots \\ s_Q(k,n) \end{pmatrix}$$

$$\boldsymbol{x}(k,n) = \boldsymbol{H}(k)\boldsymbol{s}_{kn}(k,n)$$

Examples of ATFs

- **Steering vector**
  - Under anechoic conditions, the ATF can be written as follows:

$$\tilde{\boldsymbol{h}}_q^{\mathrm{s}}(k) = \left[\frac{1}{\sqrt{4\pi} d_{1,q}} e^{-\jmath \omega_k \tau_{1,q}} \ \cdots \ \frac{1}{\sqrt{4\pi} d_{M,q}} e^{-\jmath \omega_k \tau_{M,q}} \right]^{\mathsf{T}} \quad \text{(spherical wave)}$$

$$\tilde{\boldsymbol{h}}_q^{\mathrm{p}}(k) = \alpha \left[e^{-\jmath \omega_k \tau_{1,q}} \ \cdots \ e^{-\jmath \omega_k \tau_{M,q}} \right]^{\mathsf{T}} \quad \text{(plane wave)}$$

  - f_s: sampling frequency
  - $\omega_k = 2\pi \frac{k}{K}$: normalized angular frequency
  - $\tau_{m,q} = d_{m,q} f_s / c$: propagation delay in samples

Relative Transfer Function (*RTF*)

- **Problem with the ATF: amplitude and phase ambiguity**
  - For arbitrary a and τ:

$$\begin{aligned}\boldsymbol{h}_q(k)\breve{s}_q(k,n) &= ae^{-\jmath\omega_k\tau}\boldsymbol{h}_q(k)\frac{\breve{s}_q(k,n)}{ae^{-\jmath\omega_k\tau}} \\ &= \boldsymbol{h}'_q(k)\breve{s}'_q(k,n)\end{aligned}$$

Relative Transfer Function (*RTF*)

- **Problem with the ATF: amplitude and phase ambiguity**
  - For arbitrary a and τ:

$$\begin{aligned}\boldsymbol{h}_q(k)\check{s}_q(k,n) &= ae^{-\jmath\omega_k\tau}\boldsymbol{h}_q(k)\frac{\check{s}_q(k,n)}{ae^{-\jmath\omega_k\tau}} \\ &= \boldsymbol{h}'_q(k)\check{s}'_q(k,n)\end{aligned}$$

  - Infinitely many ATF–source pairs explain the same source image
  - The original source signal cannot be identified from observations alone
  - What can be estimated is the source image
    - cf. projection back in blind source separation

Relative Transfer Function (RTF) (cont'd)

- **Definition of the RTF** [S. Doclo+, 2015]
  - r: reference microphone index
  - The ATF normalized by the transfer function at the reference microphone

$$\boldsymbol{a}_q(k) \triangleq \frac{1}{h_{r,q}(k)} \boldsymbol{h}_q(k)$$

- **Observation model based on the RTF**

$$\boldsymbol{x}(k,n) = \sum_{q=1}^{Q} \boldsymbol{a}_q(k) s_{r,q}(k,n),$$

$$s_{r,q}(k,n) = h_{r,q}(k) \check{s}_q(k,n)$$

  - $s_{r,q}(k,n)$: source image at the reference microphone

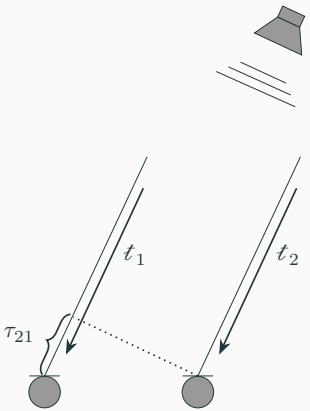
ATF and RTF in the Plane Wave Model

- **RTF between microphones 1 and 2**
  - Reference microphone: $r = 1$

$$a_2(k) = \frac{\alpha e^{-\jmath\omega_k \frac{d_{2,q}}{c} f_s}}{\alpha e^{-\jmath\omega_k \frac{d_{1,q}}{c} f_s}}$$

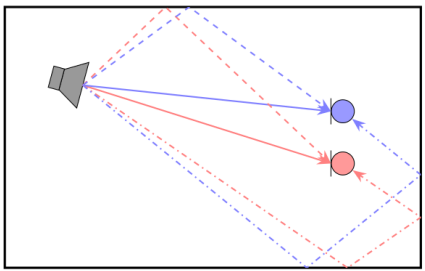
$$= e^{-\jmath\omega_k \frac{d_{2,q}-d_{1,q}}{c} f_s} = e^{-\jmath\omega_k \tau_{21} f_s}$$

  - $t_m = \frac{d_{m,q}}{c}$: Time of arrival (TOA)
  - $\tau_{21} = t_2 - t_1 = \frac{d_{2,q}-d_{1,q}}{c}$: Time difference of arrival (TDOA), or time delay
  - The RTF phase encodes the inter-microphone time delay



Summary So Far

- **Physical model of observations**
  - Sound reaches microphones through direct and reflected paths
  - Propagation is modeled as an LTI system
  - In the STFT domain, propagation is represented by an ATF or RTF



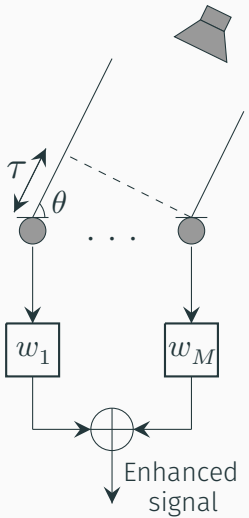
- **RTF-based observation model**
  - The RTF describes relative propagation to microphones
  - In the plane-wave model, its phase encodes the TDOA

$$\boldsymbol{x}(k,n) = \sum_{q=1}^{Q} \boldsymbol{a}_q(k) s_q(k,n)$$

Time Delay Estimation

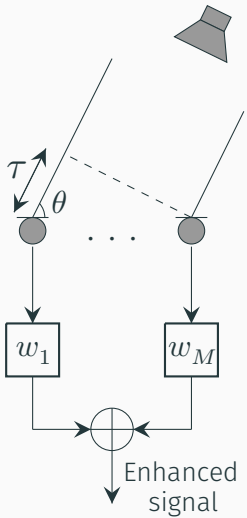
Microphone Arrays and Time Delays

- **Time delay (TD) is an important feature representing spatial information**
  - In the STFT domain, it appears as a phase difference through the RTF phase term $e^{-\jmath\omega_k\tau}$
- **TDs are used in many MASP tasks**
  - DOA estimation: $\tau = d\cos(\theta)/c$
  - Signal enhancement and source separation
  - Spatial filtering and signal synchronization



Microphone Arrays and Time Delays

- **Time delay (TD) is an important feature representing spatial information**
  - In the STFT domain, it appears as a phase difference through the RTF phase term $e^{-\jmath\omega_k\tau}$
- **TDs are used in many MASP tasks**
  - DOA estimation: $\tau = d\cos(\theta)/c$
  - Signal enhancement and source separation
  - Spatial filtering and signal synchronization
- **How should we model and estimate the TD τ?**
  - From here on, we assume a single source



Observation Models for TD Estimation [J. Chen+, 2006]

1. Ideal propagation model
2. Multipath model
3. Reverberant model

Observation Models for TD Estimation [J. Chen+, 2006]

- **Microphone observations for TD estimation can be modeled at different levels of realism**
  - From simple propagation to realistic reverberant environments

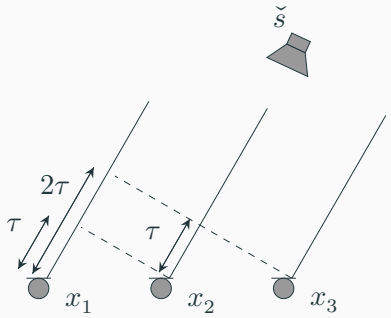
1. Ideal propagation model
2. Multipath model
3. Reverberant model

1. Ideal propagation model

- **Propagation model under the plane-wave assumption**

$$x_m(t) = \underbrace{\alpha_m}_{\text{attenuation}} \check{s}[t - \underbrace{g_m(\tau)}_{\text{TD}}] + \underbrace{u_m(t)}_{\text{noise}}$$

  - A model that considers only TDs and signal attenuation, i.e., reverberation is not considered
  - For a uniform linear array, $g_m(\tau) = (m-1)\tau$ (g_1 is always 0)
  - Goal: estimate the TD τ from $\boldsymbol{x}(t)$



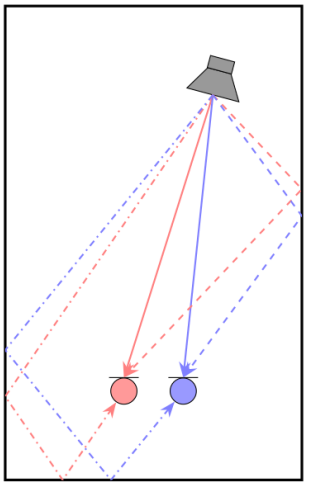
2. Multipath model

- **Model with a small number of paths**

$$x_m(t) = \sum_{p=1}^{P} \alpha_{m,p} \check{s}(t - \tau_{m,p}) + u_m(t)$$

TD of path p

  - The direct path $\tau_{m,1}$ and early reflections are modeled explicitly
  - $\alpha_{m,p}$: attenuation of the p-th path
  - Goal: estimate the delays $\tau_{m,1}$ for all m
  - Example: *underwater acoustics*, where Direct wave and reflections from the sea surface and seabed

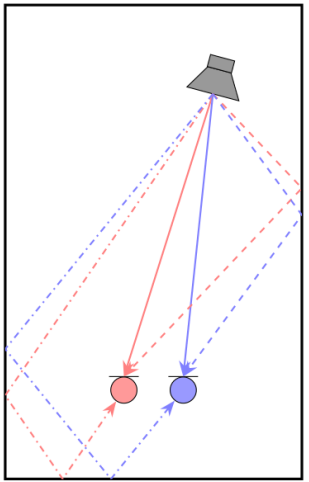


3. Reverberant model

- **Impulse-response-based model**

$$x_m(t) = h_m * \check{s}(t) + u_m(t)$$

  - The source signal is convolved with the room impulse response
  - In the STFT domain, propagation is represented by an ATF or RTF
  - The direct-path TD is not explicitly parameterized
  - Goal: 1. estimate the RTF or 2. identify the direct-path component



Cross-Correlation and TD Estimation

- **TD estimation by maximizing the weighted cross-correlation function** [C. Knapp+, 1976]
  - Widely used algorithm under the ideal propagation model
  - Estimates the TD between two microphones
  - $W_k \in \mathbb{R}_+$: arbitrary weighting function
  - For the two signals x_{1kn} and x_{2kn} (hereafter $\boldsymbol{x}(k,n)$ is written as $\boldsymbol{x}_{kn}$):

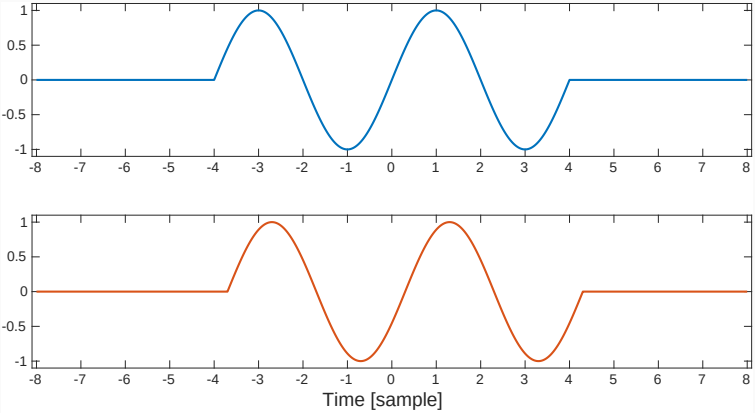
$$\hat{\tau}_{12} = \arg\max_{\tau} \ \Phi^{(12)}(\tau)$$

$$\Phi^{(12)}(\tau) = \frac{1}{K} \sum_{k=-K/2+1}^{K/2} W_k V_{12k} e^{\jmath \omega_k \tau}$$

$$V_{12k} = \frac{1}{N} \sum_{n=1}^{N} x_{1kn}^* x_{2kn}$$

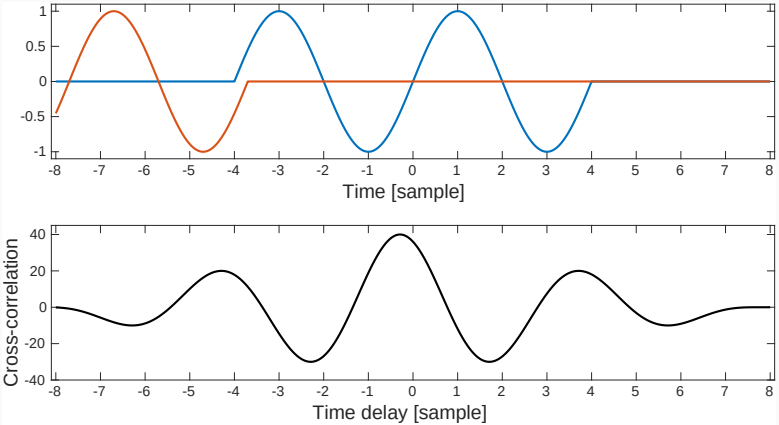
Cross-Correlation and TD estimates

- **Cross-correlation of signals $x(\tilde{t})$ and $y(\tilde{t})$:**
  $\Phi(\tau) = \int_{-\infty}^{\infty} x(\tilde{t})y(\tilde{t}+\tau)dt$



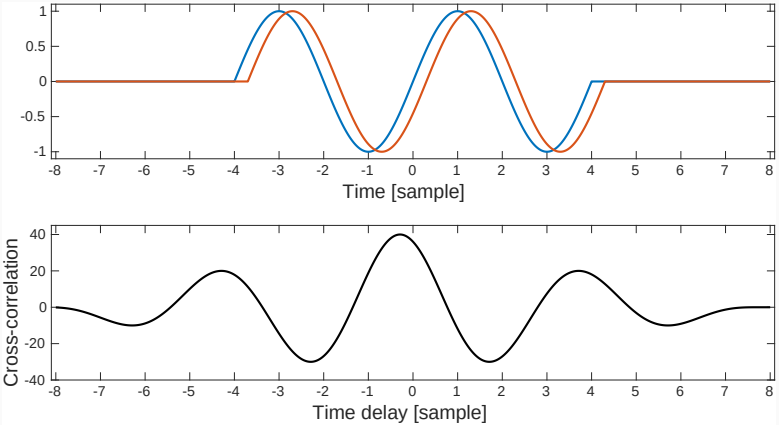
Cross-Correlation and TD estimates

- **Cross-correlation of signals $x(\tilde{t})$ and $y(\tilde{t})$:**
  $\Phi(\tau) = \int_{-\infty}^{\infty} x(\tilde{t})y(\tilde{t}+\tau)dt$



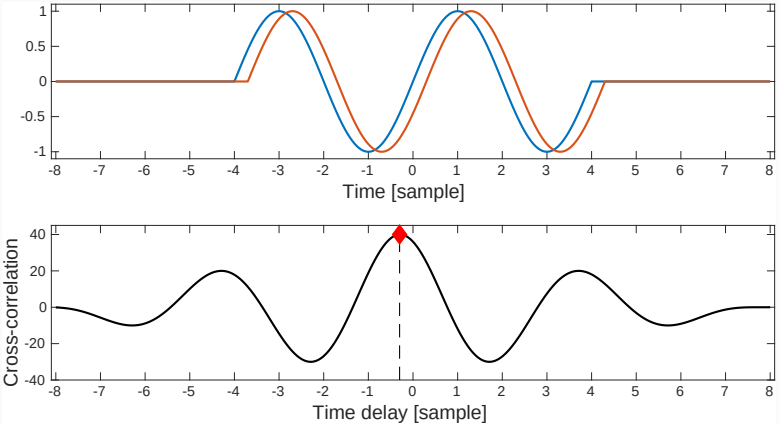
Cross-Correlation and TD estimates

- **Cross-correlation of signals $x(\tilde{t})$ and $y(\tilde{t})$:**
  $\Phi(\tau) = \int_{-\infty}^{\infty} x(\tilde{t})y(\tilde{t}+\tau)dt$



Cross-Correlation and TD estimates

- **Cross-correlation of signals $x(\tilde{t})$ and $y(\tilde{t})$:**
  $\Phi(\tau) = \int_{-\infty}^{\infty} x(\tilde{t})y(\tilde{t}+\tau)dt$



Statistical View for TD Estimation

- **Consider the ideal propagation model in the STFT domain**
  - TD τ appears as a frequency-independent phase shift

$$x_{1kn} = s_{kn} + u_{1kn}$$
$$x_{2kn} = \alpha e^{-\jmath\omega_k \tau} s_{kn} + u_{2kn}$$

- **TD estimation can be viewed as parameter estimation**
  - Parameter to be estimated: TD τ
  - Other unknowns: Source signal s_{kn}
  - For simplicity, we here assume attenuation $\alpha = 1$
- **Key question:**
  - Which value of τ best explains the observed signals?

Maximum Likelihood (ML) Estimation

- Assume additive complex Gaussian noise: $u_{mkn} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$
- ML estimation is equivalent to least-squares

$$\hat{\tau} = \underset{\tau, \{s_{kn}\}}{\arg\min} \sum_{k,n} (|x_{1kn} - s_{kn}|^2 + |x_{2kn} - s_{kn} e^{-\jmath \omega_k \tau}|^2)$$

- Step 1: Find the ML estimate of s_{kn} for a fixed τ
  - Setting the Wirtinger derivative to zero gives

$$\hat{s}_{kn}(\tau) = \frac{1}{2}(x_{1kn} + x_{2kn} e^{\jmath \omega_k \tau})$$

Maximum Likelihood (ML) Estimation

- Step 2: Substitute $\hat{s}_{kn}(\tau)$ into the objective function and solve the following least-squares:

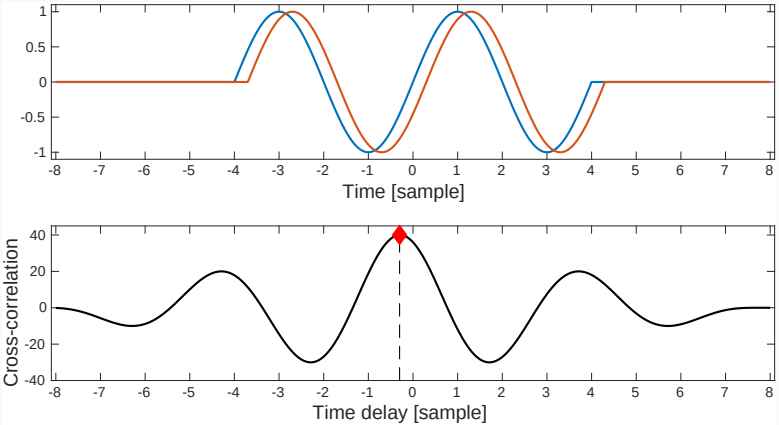
$$\hat{\tau} = \arg\max_{\tau} \frac{1}{2} \sum_{k,n} |x_{1kn} - x_{2kn} e^{\jmath \omega_k \tau}|^2$$

$$= \arg\max_{\tau} \Re\left\{ \underbrace{\sum_{k,n} x_{1kn}^* x_{2kn} e^{\jmath \omega_k \tau}}_{\text{Cross-correlation}} \right\}$$

- This has the same form as cross-correlation-based TD estimation

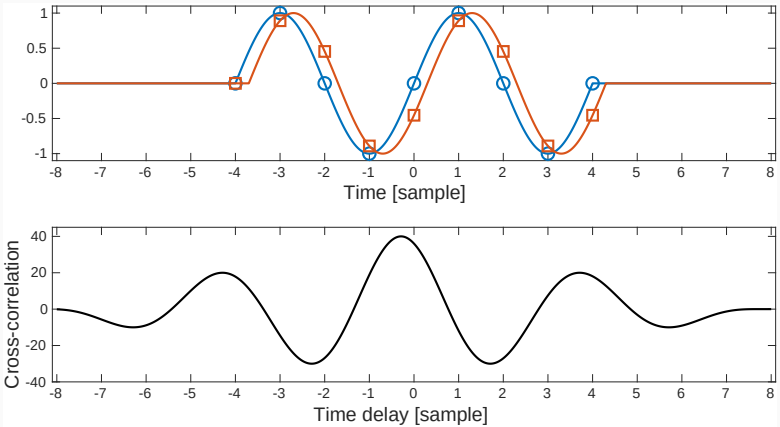
TD Estimation by Maximizing Cross-Correlation

- Note: can we just find the TD corresponding to the maximum value directly?



Cross-Correlation and TD for Discrete Signals

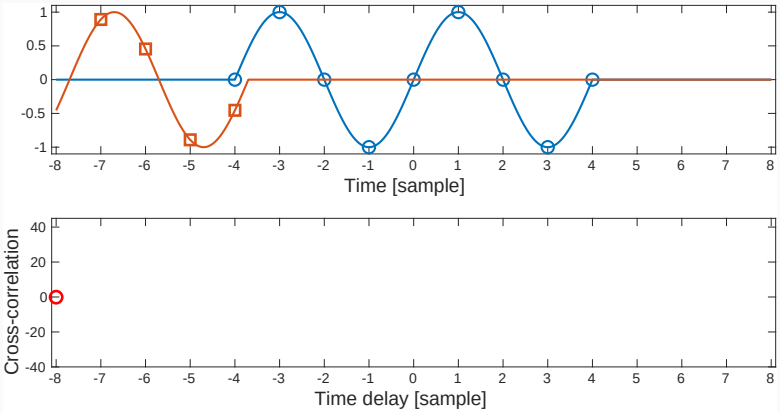
- **Cross-correlation of discrete signals $x[n]$ and $y[n]$:**
  $\Phi[\nu] = \sum_{n=-\infty}^{\infty} x[n]y[n+\nu]$



Cross-Correlation and TD for Discrete Signals

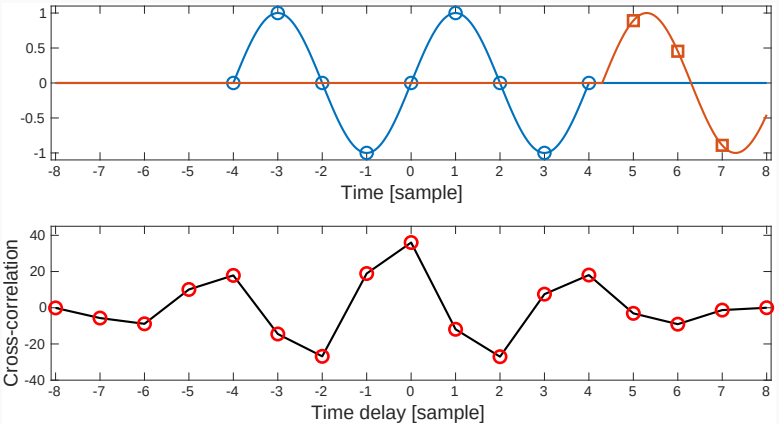
- **Cross-correlation of discrete signals $x[n]$ and $y[n]$:**

$$\Phi[\nu] = \sum_{n=-\infty}^{\infty} x[n]y[n+\nu]$$



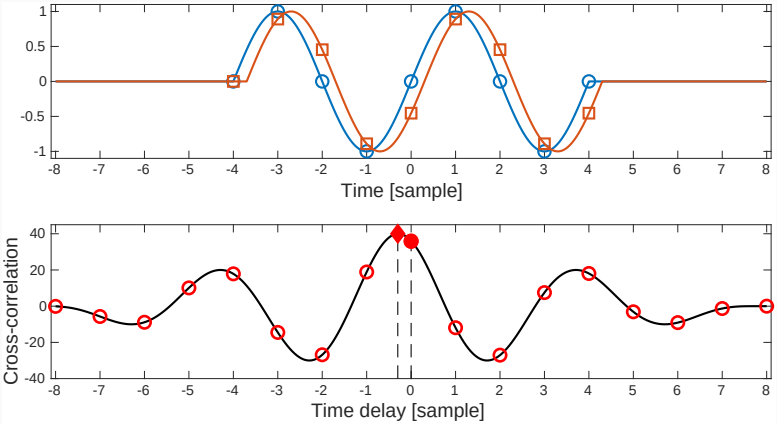
Cross-Correlation and TD for Discrete Signals

- **Cross-correlation of discrete signals $x[n]$ and $y[n]$:**
  $\Phi[\nu] = \sum_{n=-\infty}^{\infty} x[n]y[n+\nu]$



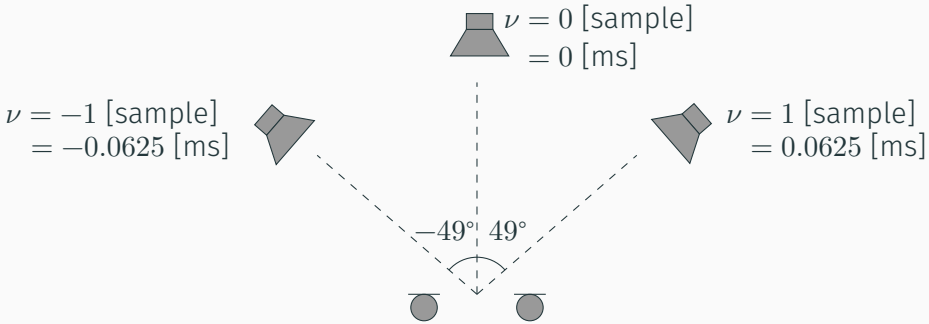
Discrete TD estimate

- Simple maximization has a limited estimation accuracy



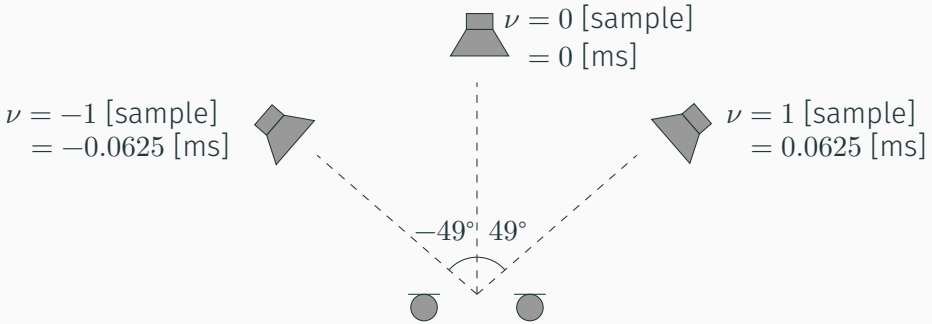
Example of Simple Estimation

- **In general, cross-correlation is computed for discrete delays ν**
  - Example: microphone spacing 3.4 cm, sampling frequency 16 kHz, speed of sound 340 m/s



Example of Simple Estimation

- **In general, cross-correlation is computed for discrete delays ν**
  - Example: microphone spacing 3.4 cm, sampling frequency 16 kHz, speed of sound 340 m/s
  - Estimation at subsample is required

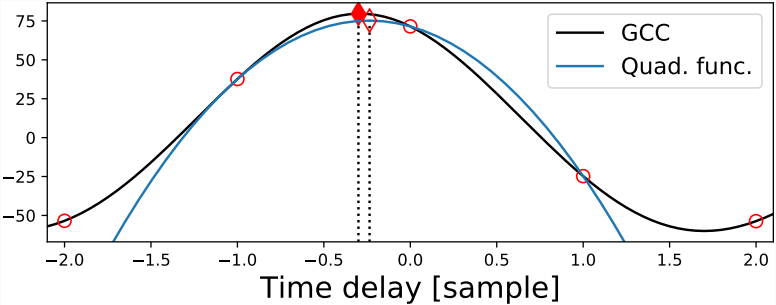


Various Sub-Sample TD Estimation Methods

- **Interpolation-based methods**
  - Quadratic interpolation [Jacovitti+, 1993]
  - Gaussian fitting [Zhang+, 2006]
  - Zero-padding in the frequency domain
- **Search-based methods**
  - Ternary search, golden section search [Kiefer, 1953]
- **Analytical methods**
  - Auxiliary-function-based method [Yamaoka+, 2019]

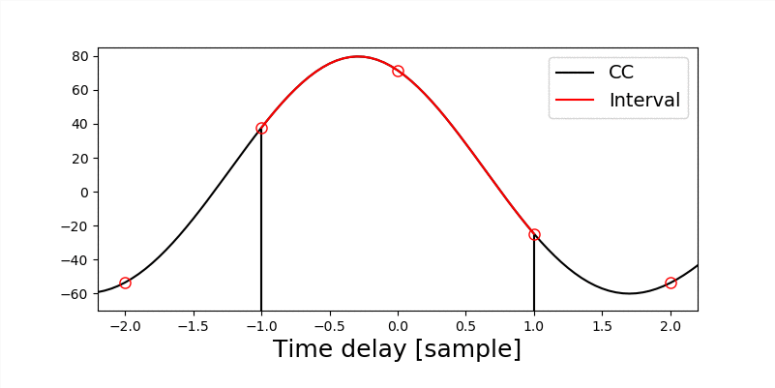
Quadratic Interpolation

- **Interpolate the cross-correlation around the discrete maximum**
  - Approximate solution using quadratic interpolation



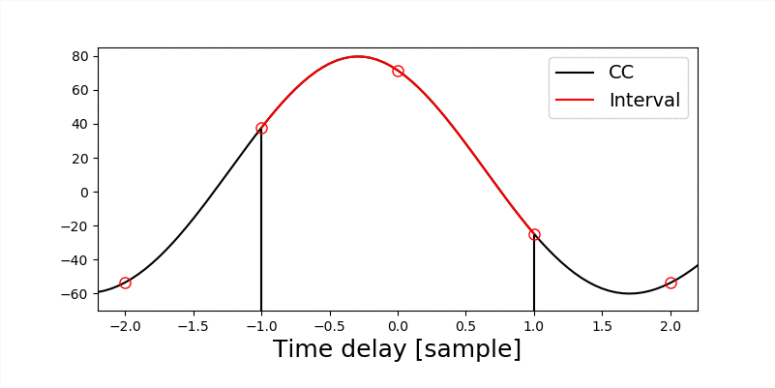
Search-Based Estimation

- **Golden section search (GSS)**
  - Searches for the maximum of a unimodal function by narrowing the search interval



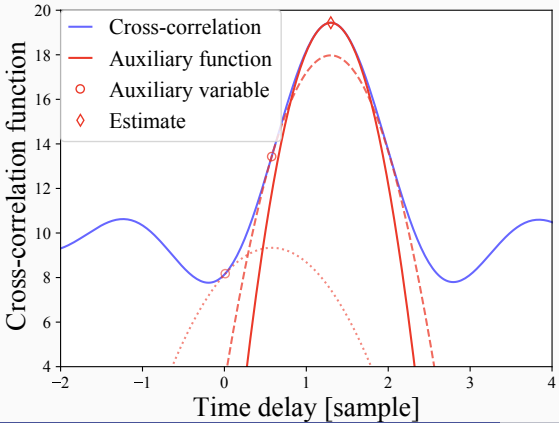
Search-Based Estimation

- **Golden section search (GSS)**
  - Searches for the maximum of a unimodal function by narrowing the search interval



Auxiliary-Function-Based Estimation

- **AuxTDE (auxiliary-function-based time delay estimation)**
  - Estimates the maximum of the cross-correlation by iteratively maximizing an auxiliary function



Evaluation of TD Estimation

- TD estimation has been formulated by two key components
  - Observation model
  - Optimization algorithm
- The next question is evaluation
  - How accurate can TD estimation be?
  - What determines the estimation accuracy?

Evaluation of TD Estimation

- TD estimation has been formulated by two key components
  - Observation model
  - Optimization algorithm
- The next question is evaluation
  - How accurate can TD estimation be?
  - What determines the estimation accuracy?
- In addition to experiments, theoretical analysis is often used in signal processing

▶ Example: Cramér–Rao lower bound (CRLB)
  - A lower bound on the variance of unbiased estimators

Cramér–Rao Lower Bound (1/2)

- **Setup**
  - Observation: $\boldsymbol{x}_{kn}$, Unknown parameter: $\boldsymbol{\theta}$
  - Unbiased estimator: $\boldsymbol{T}(\boldsymbol{x}_{kn})$ satisfying $\mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})] = \boldsymbol{\theta}$
  - Evaluation criterion: mean squared error (MSE)

- **Q. What is the best unbiased estimator in terms of MSE?**

$$\begin{aligned}\mathrm{MSE} &= \mathbb{E}\left[\{\boldsymbol{T}(\boldsymbol{x}_{kn}) - \boldsymbol{\theta}\}^2\right] \\ &= \mathbb{E}\left[\{\boldsymbol{T}(\boldsymbol{x}_{kn}) - \mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})]\}^2\right] + \{\mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})] - \boldsymbol{\theta}\}^2 \qquad (1)\end{aligned}$$

Cramér–Rao Lower Bound (1/2)

- **Setup**
  - Observation: $\boldsymbol{x}_{kn}$, Unknown parameter: $\boldsymbol{\theta}$
  - Unbiased estimator: $\boldsymbol{T}(\boldsymbol{x}_{kn})$ satisfying $\mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})] = \boldsymbol{\theta}$
  - Evaluation criterion: mean squared error (MSE)

- **Q. What is the best unbiased estimator in terms of MSE?**

$$\begin{aligned}\mathrm{MSE} &= \mathbb{E}\left[\{\boldsymbol{T}(\boldsymbol{x}_{kn}) - \boldsymbol{\theta}\}^2\right] \\ &= \mathbb{E}\left[\{\boldsymbol{T}(\boldsymbol{x}_{kn}) - \mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})]\}^2\right] + \{\mathbb{E}[\boldsymbol{T}(\boldsymbol{x}_{kn})] - \boldsymbol{\theta}\}^2 \qquad (1)\end{aligned}$$

Variance → should be minimized

Bias $= 0$

Cramér–Rao Lower Bound (2/2)

- **Statement of the CRLB**
  - For an unknown parameter vector $\boldsymbol{\theta} = [\theta_1 \ \cdots \ \theta_M]$, any unbiased estimator $\tilde{\boldsymbol{\theta}}$ satisfies

$$\operatorname{cov}(\tilde{\boldsymbol{\theta}}) \geq \mathcal{I}^{-1}(\boldsymbol{\theta}) \tag{2}$$

  - This gives a theoretical performance limit for unbiased estimators
  - $\mathcal{I}(\boldsymbol{\theta})$: Fisher information matrix (FIM)

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \mathcal{L}(\boldsymbol{\theta}|\boldsymbol{x}_{kn})\right] \tag{3}$$

Preparation

- Observation model for an M-channel microphone array

$$\boldsymbol{x}_{kn} = s_{kn}\boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) + \boldsymbol{u}_{kn} \quad \text{(observation)} \tag{4}$$

$$\boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) = [a_1 e^{-\jmath\omega_k\tau_{r1}} \ \cdots \ a_M e^{-\jmath\omega_k\tau_{rM}}]^{\mathsf{T}} \quad \text{(ideal propagation)} \tag{5}$$

$$\boldsymbol{\tau} = [\tau_{r1} \ \cdots \ \tau_{rM}]^{\mathsf{T}} \quad \text{(TD vector)} \tag{6}$$

- By definition, $\tau_{rr} = 0$ and $\tau_{ij} = -\tau_{ji}$
- Problem: estimate $\bar{\boldsymbol{\tau}}_r$ under the constraint $\tau_{rr} = 0$
  - $\bar{\boldsymbol{\tau}}_r$: $(M-1)$-dimensional TD vector obtained by removing τ_{rr} from $\boldsymbol{\tau}$

Likelihood Function

- **Assumptions**
  - Target source: $s_{kn} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_{\mathrm{s}}^2)$, Noise: $\boldsymbol{u}_{kn} \sim \mathcal{N}_{\mathbb{C}}(0, \boldsymbol{\Sigma}_{\mathrm{u}})$
  - Noise covariance: $\boldsymbol{\Sigma}_{\mathrm{u}} = \mathrm{diag}([\sigma_1^2 \ \cdots \ \sigma_M^2])$
  - Known parameters: σ_{s}^2, $\boldsymbol{\Sigma}_{\mathrm{u}}$, and a_m $(m = 1, \ldots, M)$
- **By the reproductive property, $\boldsymbol{x}_{kn}$ follows $\mathcal{N}_{\mathbb{C}}(0, \boldsymbol{R}_k(\boldsymbol{\tau}))$**

$$\boldsymbol{R}_k(\boldsymbol{\tau}) = \sigma_{\mathrm{s}}^2 \boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) \boldsymbol{a}_k^{\mathsf{H}}(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) + \boldsymbol{\Sigma}_{\mathrm{u}} \tag{7}$$

- **Likelihood function of $\boldsymbol{\tau}$**

$$\mathcal{L}(\boldsymbol{\tau} \mid \boldsymbol{x}_{kn}) = \prod_{k,n} \frac{1}{\pi^M \det(\boldsymbol{R}_k(\boldsymbol{\tau}))} \exp\{-\boldsymbol{x}_{kn}^{\mathsf{H}} \boldsymbol{R}_k^{-1}(\boldsymbol{\tau}) \boldsymbol{x}_{kn}\} \tag{8}$$

Log-Likelihood Function (1/2)

- The log-likelihood function is given by

$$\log \mathcal{L}(\boldsymbol{\tau} \mid \boldsymbol{x}_{kn}) = -\sum_{k,n} \{\log \pi^M + \log \det(\boldsymbol{R}_k(\boldsymbol{\tau})) + \boldsymbol{x}_{kn}^{\mathsf{H}} \boldsymbol{R}_k^{-1}(\boldsymbol{\tau}) \boldsymbol{x}_{kn}\} \quad (9)$$

- Dependence on $\boldsymbol{\tau}$
  - First term: constant
  - Second term: constant by Sylvester's determinant theorem (Weinstein–Aronszajn identity)
  - Third term: simplified using the matrix inversion lemma (Sherman-Morrison-Woodbury)

Log-Likelihood Function (2/2)

- The log-likelihood function is reduced to

$$\log \mathcal{L}(\boldsymbol{\tau} \mid \boldsymbol{x}_{kn}) = \frac{\sigma_{\mathrm{s}}^2}{1 + \Lambda_M} \sum_k \tilde{\boldsymbol{g}}_k^{\mathsf{H}}(\boldsymbol{\tau}) \boldsymbol{V}_k \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) + \mathrm{C} \tag{9}$$

$$\boldsymbol{V}_k = \sum_n \boldsymbol{x}_{kn} \boldsymbol{x}_{kn}^{\mathsf{H}}, \qquad \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) = \boldsymbol{\Sigma}_{\mathrm{u}}^{-1} \boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) \tag{10}$$

$$\Lambda_M = \sum_{m=1}^{M} \lambda_m, \qquad \lambda_m = \frac{a_m^2 \sigma_{\mathrm{s}}^2}{\sigma_m^2} \tag{11}$$

- Interpretation
  - The likelihood is a weighted sum of cross-correlations over all microphone pairs
  - The likelihood depends on channel-wise SNRs λ_m

Log-Likelihood Function (2/2)

- **The log-likelihood function is reduced to**

Corresponds to a sum of cross-correlations

$$\log \mathcal{L}(\boldsymbol{\tau} \mid \boldsymbol{x}_{kn}) = \frac{\sigma_{\mathrm{s}}^2}{1 + \Lambda_M} \sum_k \tilde{\boldsymbol{g}}_k^{\mathsf{H}}(\boldsymbol{\tau}) \boldsymbol{V}_k \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) + \mathrm{C} \tag{9}$$

$$\boldsymbol{V}_k = \sum_n \boldsymbol{x}_{kn} \boldsymbol{x}_{kn}^{\mathsf{H}}, \qquad \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) = \boldsymbol{\Sigma}_{\mathrm{u}}^{-1} \boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) \tag{10}$$

$$\Lambda_M = \sum_{m=1}^{M} \lambda_m, \qquad \lambda_m = \frac{a_m^2 \sigma_{\mathrm{s}}^2}{\sigma_m^2} \tag{11}$$

- **Interpretation**
  - The likelihood is a weighted sum of cross-correlations over all microphone pairs
  - The likelihood depends on channel-wise SNRs λ_m

Log-Likelihood Function (2/2)

- **The log-likelihood function is reduced to**

Corresponds to a sum of cross-correlations

$$\log \mathcal{L}(\boldsymbol{\tau} \mid \boldsymbol{x}_{kn}) = \frac{\sigma_{\mathrm{s}}^2}{1 + \Lambda_M} \sum_k \tilde{\boldsymbol{g}}_k^{\mathsf{H}}(\boldsymbol{\tau}) \boldsymbol{V}_k \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) + \mathrm{C} \tag{9}$$

$$\boldsymbol{V}_k = \sum_n \boldsymbol{x}_{kn} \boldsymbol{x}_{kn}^{\mathsf{H}}, \qquad \tilde{\boldsymbol{g}}_k(\boldsymbol{\tau}) = \boldsymbol{\Sigma}_{\mathrm{u}}^{-1} \boldsymbol{a}_k(\boldsymbol{\alpha}_k, \boldsymbol{\tau}) \tag{10}$$

$$\Lambda_M = \sum_{m=1}^{M} \lambda_m, \qquad \lambda_m = \frac{a_m^2 \sigma_{\mathrm{s}}^2}{\sigma_m^2} \tag{11}$$

SNR at channel m

- **Interpretation**
  - The likelihood is a weighted sum of cross-correlations over all microphone pairs
  - The likelihood depends on channel-wise SNRs λ_m

FIM for TDs (1/2)

- **Each element of the FIM**
  - Obtained by differentiating the log-likelihood and taking the expectation

$$[\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r)]_{ii} = \frac{2}{1+\Lambda_M} \sum_{k,n} \omega_k^2 \sum_{m \neq i} \lambda_i \lambda_m \tag{12a}$$

$$[\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r)]_{ij} = -\frac{2}{1+\Lambda_M} \sum_{k,n} \omega_k^2 \lambda_i \lambda_j \tag{12b}$$

FIM for TDs (1/2)

- **Each element of the FIM**
  - Obtained by differentiating the log-likelihood and taking the expectation

$$\left[\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r)\right]_{ii} = \frac{2}{1+\Lambda_M}\sum_{k,n}\omega_k^2\sum_{m\neq i}\lambda_i\lambda_m \tag{12a}$$

$$\left[\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r)\right]_{ij} = -\frac{2}{1+\Lambda_M}\sum_{k,n}\omega_k^2\lambda_i\lambda_j \tag{12b}$$

- **Interpretation**
  - Diagonal elements depend on the SNR products involving channel i
  - Off-diagonal elements represent coupling between channels i and j

FIM for TDs (2/2)

- **FIM:** $\mathcal{I}_r : \mathbb{R}^{M-1} \to \mathbb{R}^{(M-1)\times(M-1)}$

$$\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r) = \frac{2}{1+\Lambda_M} \sum_{k,n} \omega_k^2 \left[\Lambda_M \mathrm{diag}(\bar{\boldsymbol{\lambda}}) - \bar{\boldsymbol{\lambda}}\bar{\boldsymbol{\lambda}}^{\mathsf{T}}\right] \tag{12}$$

$$\bar{\boldsymbol{\lambda}} = \left[\lambda_1 \ \cdots \ \lambda_{r-1} \ \lambda_{r+1} \ \cdots \ \lambda_M\right]^{\mathsf{T}} \tag{13}$$

  - This result agrees with [W. Hahn+, 1973], except for the inclusion of signal attenuation a_m

FIM for TDs (2/2)

- **FIM:** $\mathcal{I}_r : \mathbb{R}^{M-1} \to \mathbb{R}^{(M-1)\times(M-1)}$

$$\mathcal{I}_r(\bar{\boldsymbol{\tau}}_r) = \frac{2}{1+\Lambda_M} \sum_{k,n} \omega_k^2 \left[\Lambda_M \mathrm{diag}(\bar{\boldsymbol{\lambda}}) - \bar{\boldsymbol{\lambda}}\bar{\boldsymbol{\lambda}}^{\mathsf{T}}\right] \tag{12}$$

$$\bar{\boldsymbol{\lambda}} = [\lambda_1 \ \cdots \ \lambda_{r-1} \ \lambda_{r+1} \ \cdots \ \lambda_M]^{\mathsf{T}} \tag{13}$$

  - This result agrees with [W. Hahn+, 1973], except for the inclusion of signal attenuation a_m
  - The matrix inversion lemma can be applied by exploiting this structure

Closed-Form CRLB

- The CRLB for TD estimation is given by

$$\mathcal{I}_r^{-1}(\bar{\boldsymbol{\tau}}_r) = \beta_M \left(\mathrm{diag}(\bar{\boldsymbol{\lambda}})^{-1} + \frac{1}{\lambda_r} 1 1^{\mathsf{T}} \right) \tag{14}$$

$$\beta_M = \left(\frac{2\Lambda_M}{1+\Lambda_M} \sum_{k,n} \omega_k^2 \right)^{-1} \tag{15}$$

  - Under $M = 2$ and $\lambda_1 = \lambda_2$, this expression agrees with the conventional CRLB for cross-correlation-based TD estimation [C. Knapp+, 1976]

Example of CRLB ($M = 4, r = 1$)

- Each variance lower bound depends on the SNRs at the reference and target microphones, while the overall scale is determined by β_M

$$\begin{pmatrix} \mathrm{var}(\hat{\tau}_{12}) & \mathrm{cov}(\hat{\tau}_{12}, \hat{\tau}_{13}) & \mathrm{cov}(\hat{\tau}_{12}, \hat{\tau}_{14}) \\ \mathrm{cov}(\hat{\tau}_{13}, \hat{\tau}_{12}) & \mathrm{var}(\hat{\tau}_{13}) & \mathrm{cov}(\hat{\tau}_{13}, \hat{\tau}_{14}) \\ \mathrm{cov}(\hat{\tau}_{14}, \hat{\tau}_{12}) & \mathrm{cov}(\hat{\tau}_{14}, \hat{\tau}_{13}) & \mathrm{var}(\hat{\tau}_{14}) \end{pmatrix} \geq \beta_4 \begin{pmatrix} \frac{1}{\lambda_1} + \frac{1}{\lambda_2} & \frac{1}{\lambda_1} & \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \frac{1}{\lambda_1} + \frac{1}{\lambda_3} & \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \frac{1}{\lambda_1} & \frac{1}{\lambda_1} + \frac{1}{\lambda_4} \end{pmatrix}$$

$$\beta_4 = \left(\frac{2(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)}{1 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \sum_{k,n} \omega_k^2 \right)^{-1}$$

Example of CRLB ($M = 4, r = 1$)

- Each variance lower bound depends on the SNRs at the reference and target microphones, while the overall scale is determined by β_M

$$\begin{pmatrix} \mathrm{var}(\hat{\tau}_{12}) & \mathrm{cov}(\hat{\tau}_{12}, \hat{\tau}_{13}) & \mathrm{cov}(\hat{\tau}_{12}, \hat{\tau}_{14}) \\ \mathrm{cov}(\hat{\tau}_{13}, \hat{\tau}_{12}) & \mathrm{var}(\hat{\tau}_{13}) & \mathrm{cov}(\hat{\tau}_{13}, \hat{\tau}_{14}) \\ \mathrm{cov}(\hat{\tau}_{14}, \hat{\tau}_{12}) & \mathrm{cov}(\hat{\tau}_{14}, \hat{\tau}_{13}) & \mathrm{var}(\hat{\tau}_{14}) \end{pmatrix} \geq \beta_4 \begin{pmatrix} \frac{1}{\lambda_1} + \frac{1}{\lambda_2} & \frac{1}{\lambda_1} & \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \frac{1}{\lambda_1} + \frac{1}{\lambda_3} & \frac{1}{\lambda_1} \\ \frac{1}{\lambda_1} & \frac{1}{\lambda_1} & \frac{1}{\lambda_1} + \frac{1}{\lambda_4} \end{pmatrix}$$

$$\beta_4 = \left(\frac{2(\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4)}{1 + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \sum_{k,n} \omega_k^2 \right)^{-1}$$

Interpretation of the CRLB

- **The CRLB for τ_{rm} is determined by two factors:**

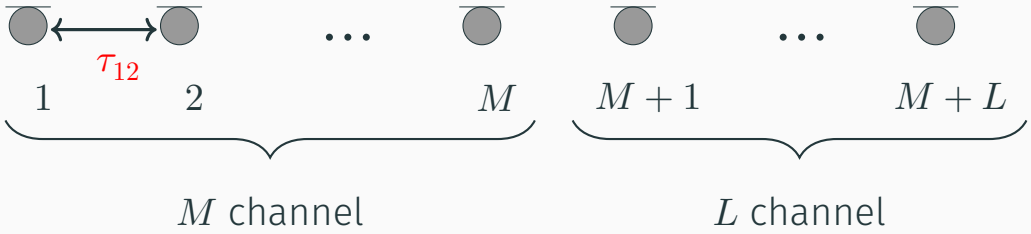
$$\mathrm{var}(\hat{\tau}_{rm}) \geq \beta_M \left(\frac{1}{\lambda_r} + \frac{1}{\lambda_m} \right) \tag{16}$$

$$\beta_M = \frac{1 + \Lambda_M}{2\Lambda_M \sum_{k,n} \omega_k^2} = \frac{1 + 1/\Lambda_M}{2 \sum_{k,n} \omega_k^2} \tag{17}$$

  - Larger total SNR Λ_M decreases the overall scale β_M, although the improvement saturates at high SNR
  - Larger SNRs at the microphone pair (r, m) decrease the variance bound for τ_{rm}

Do More Microphones Help?

- **Target: estimate τ_{12} from microphone-array observations**
- **Question:**
  - How much improvement is obtained by adding L microphones to an M-microphone array?



Number of Sensors and CRLB

- For a given TD, define the CRLB improvement factor as

$$\zeta = \frac{\mathrm{CRLB}_M}{\mathrm{CRLB}_{M+L}} = \left(1 + \frac{1}{\Lambda_M}\right)\left(1 + \frac{1}{\Lambda_{M+L}}\right)^{-1} \tag{18}$$

- Since $\Lambda_{M+L} \geq \Lambda_M$, we have $\zeta \geq 1$
- More microphones improve the CRLB if computational cost is ignored
  - Q. What is the limit of this improvement?

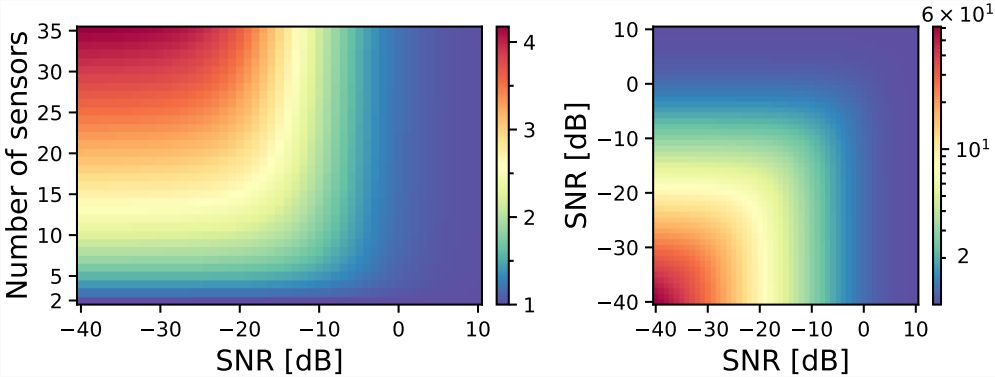
Condition for Reaching the Improvement Limit

- **Limit with respect to the SNR of microphone m**

$$\lim_{\lambda_m \to \infty} \zeta = \begin{cases} 1 & (1 \leq m \leq M) \\ 1 + 1/\Lambda_M & (M < m \leq M + L) \end{cases} \tag{19}$$

  - Existing high-SNR microphones do not increase the improvement factor due to additional microphones
  - An additional high-SNR microphone can reach the maximum improvement
  - ▶ One sufficiently high-SNR additional microphone is enough

Value of ζ with Respect to SNR



- The numerical results confirm that adding microphones is more beneficial when the original SNR is low

From Theory to Numerical Evaluation

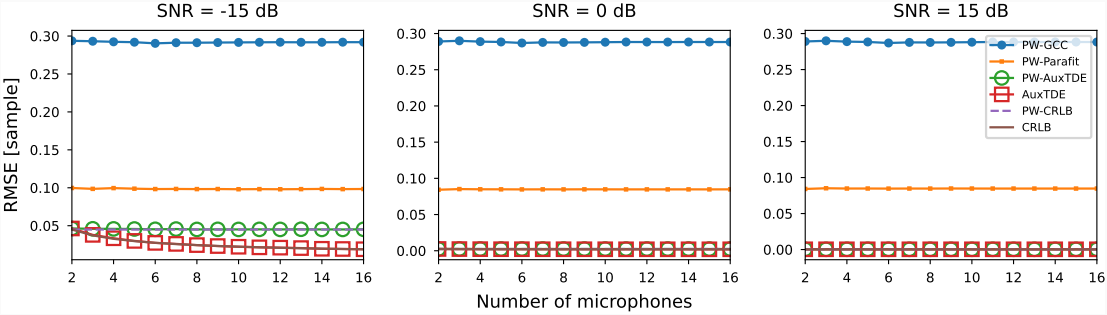
- So far, TD estimation has been discussed from three viewpoints
  - Observation model: how TDs appear in microphone signals
  - Optimization: how TDs can be estimated from observations
  - Theoretical analysis: what performance limit is given by the CRLB
- The remaining question is numerical evaluation
  - How well do practical TD estimation algorithms perform?
  - Can they approach the CRLB under ideal conditions?

▶ Next, numerical experiments are used to compare practical algorithms with the theoretical performance limit

Numerical Experiments

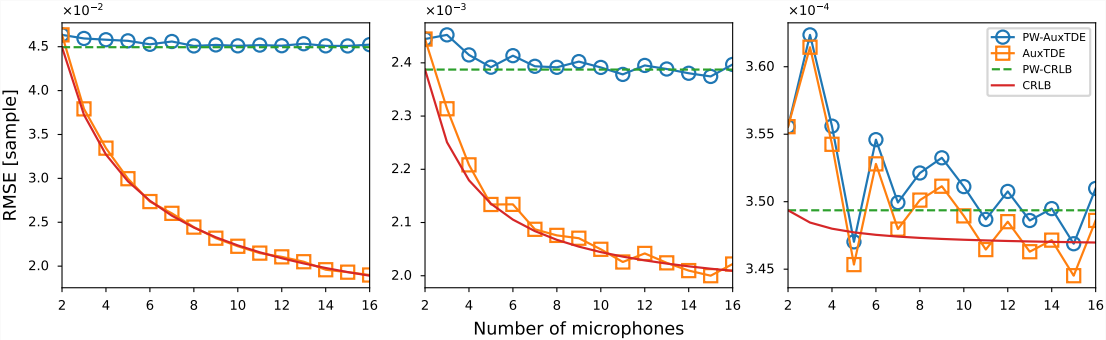
- **Experimental setup**
  - Room simulation using pyroomacoustics [Scheibler+, 2018]: $18\,\mathrm{m} \times 24\,\mathrm{m} \times 6\,\mathrm{m}$, 10000 random trials
  - Observation generated by convolving white noise with impulse responses
  - Random placement of a microphone array and a single source
- **Compared methods**
  - PW-GCC: pairwise GCC without interpolation
  - PW-Parafit: pairwise GCC with parabolic interpolation
  - PW-AuxTDE: pairwise AuxTDE [K. Yamaoka+, 2019]
  - AuxTDE: proposed method [K. Yamaoka+, 2021]
  - PW-CRLB / CRLB: pairwise and global CRLBs

Effect of the Number of Microphones (1/2)



- Discrete maximization (PW-GCC) and parabolic interpolation (PW-Parafit) do not provide sufficient estimation accuracy
- AuxTDE, which can find the maximum of the cross-correlation function, achieves performance close to the CRLB

Effect of the Number of Microphones (2/2)



- Pairwise methods do not benefit from additional microphones
- AuxTDE approaches the CRLB
- More microphones are especially effective in low-SNR conditions

Conclusion

Summary (1/2)

- Array signal processing exploits spatial information
  - Sound waves propagate through space
  - Observations depend on source and sensor positions
- Microphone observations are modeled by sound propagation
  - Direct and reflected paths are modeled as an LTI system
  - In the STFT domain, propagation is represented by an ATF or RTF
- The RTF connects propagation to TD estimation
  - It describes relative propagation to microphones
  - Under the plane-wave model, its phase encodes the TD

▶ TD estimation is a basic tool for exploiting spatial information

Summary (2/2)

- **TD estimation was discussed from three viewpoints**
  - Observation models: ideal, multipath, and reverberant models
  - Optimization: cross-correlation and its ML interpretation
  - Analysis: CRLB as a theoretical performance limit
  - Numerical experiments: validation of algorithm performance and theoretical insights
- **CRLB analysis clarifies the factors affecting accuracy**
  - Higher SNRs/Additional microphones reduce the variance bound
  - Numerical experiments validated the theoretical insights

▶ **Accurate TD estimation supports reliable spatial acoustic signal processing**

Where Does TD Estimation Appear Today?

- **Robust sound source localization** [P.-A. Grumiaux+, 2022]
  - TD estimation under noise, reverberation, and multiple sources
- **Distributed microphone arrays** [M. Cobos+, 2017]
  - Spatial processing with many microphones over a large area
  - Synchronization and calibration are central issues
- **Learning-based spatial audio processing** [R. Haeb-Umbach+, 2024]
  - Neural networks are increasingly used for localization and enhancement
  - TDs and spatial features still provide important physical structure

▶ **The theory introduced here provides a basis for understanding these systems**