Source Localisation and Blind Source Separation (UDRC Summer School)

Course Lecture Notes and Tutorial Questions

Dr James R. Hopgood

Source Localisation and Blind Source Separation (UDRC Summer School)

Course Lecture Notes and Tutorial Questions

Dr James R. Hopgood
Room 2.05
Alexander Graham Bell Building
The King's Buildings
Mayfield Road
Edinburgh
EH9 3JL
Scotland, UK
James.Hopgood@ed.ac.uk
Telephone: +44 (0)131 650 5571
Fax: +44 (0)131 650 6554

Last revision: June, 2014

School of Engineering
College of Science and Engineering
University of Edinburgh

Room 2.05
Alexander Graham Bell Building
The King's Buildings
Mayfield Road
Edinburgh
EH9 3JL
Scotland, UK
James.Hopgood@ed.ac.uk
Telephone: +44 (0)131 650 5571
Fax: +44 (0)131 650 6554.

Major revision, June, 2014.
Last printed revision with minor corrections, 14 July, 2014.

Typeset by the author with the LaTeX 2_ε Documentation System, with $\mathcal{AMS}$-LaTeX Extensions, in 12/18 pt Times and Euler fonts.

INSTITUTE FOR DIGITAL COMMUNICATIONS,
School of Engineering,
College of Science and Engineering,
Kings's Buildings,
Edinburgh, EH9 3JL. U.K.

Copyright Statement

This document does not contain copyright material.

The author of this document

1. holds the copyright for all lecture and course materials in this module;
2. holds the copyright for students notes, summaries, or recordings that substantially reflect the lecture content or materials;
3. makes these materials available only for personal use by students studying this module;
4. reserves the right that no part of the notes, tutorials, solutions, or other course materials may be distributed or reproduced for commercial purposes without express written consent from the author; this does not prevent students from sharing notes on an individual basis for personal use.

These lecture notes consist of entirely original work, where all material has been written and typeset by the author. No figures or substantial pieces of text has been reproduced verbatim from other texts.

However, there is some material that has been based on work in a number of previous textbooks, and therefore some sections and paragraphs have strong similarities in structure and wording. These texts have been referenced and include, amongst a number of others, in order of contributions:

- Huang Y., J. Benesty, and J. Chen, "Time Delay Estimation and Source Localization," in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.

  DiBiase J. H., H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, , Springer Berlin Heidelberg, 2001.

Outline of Lecture Contents

Contents

List of Figures

Acronyms

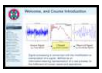
2-D	two-dimensional
AED	adaptive eigenvalue decomposition
AIR	acoustic impulse response
ASL	acoustic source localisation
AVS	acoustic vector sensor
BSS	blind source separation
CPSD	cross-power spectral density
DUET	degenerate unmixing estimation technique
FT	Fourier transform
GCC	generalised cross correlation
GCC-PHAT	GCC-PHAT
LI	linear intersection
LS	least-squares
LSE	least-squares estimate
LSE	least squares error
ML	maximum-likelihood
PHAT	phase transform
PHD	Ph.D. thesis
RIR	room impulse response
SBF	steered beamformer
SBS	block stationary
SCOT	Smoothed Coherence Transform

SI	spherical interpolation
SRC	stochastic region contraction
SRP	steered response power
STFT	short-time Fourier transform
SX	spherical intersection
TDOA	time-difference of arrival
TF	time-frequency
TFR	time-frequency representation
ULA	uniform linear array
WDO	W-disjoint orthogonality
i. t. o.	in terms of

1 Source Localisation

This tutorial looks at the role of acoustic source localisation (ASL) in block stationary (SBS), as well as how blind source separation (BSS) can be used in ASL.

1.1 Introduction



New slide

- This research tutorial is intended to cover a wide range of aspects which link acoustic source localisation (ASL) and blind source separation (BSS). It is written at a level which assumes knowledge of undergraduate mathematics and signal processing nomenclature, but otherwise should be accessible to most technical graduates.

KEYPOINT! (Latest Slides). Please note the following:

- This tutorial is being continually updated, and feedback is welcomed. The documents published on the USB stick may differ to the slides presented on the day. In particular, there are likely to be a few typos in the document, so if there is something that isn't clear, please feel free to email me so I can correct it (or make it clearer).
- The latest version of this document can be found online and downloaded at:

  `http://www.see.ed.ac.uk/~jhopgood1/Research/UDRC`

- Thanks to Xionghu Zhong and Ashley Hughes for borrowing some of their diagrams from their dissertations.

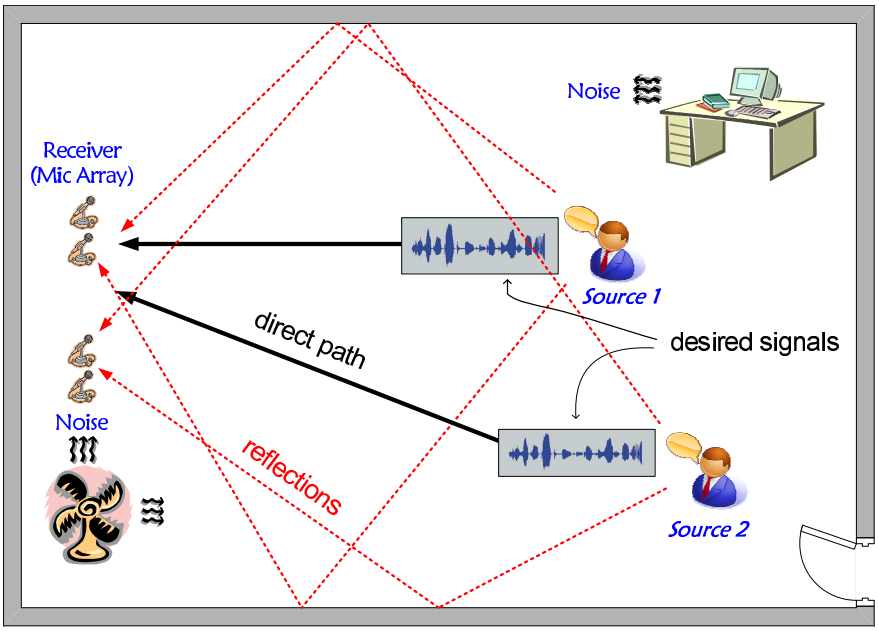


Figure 1.1: Source localisation and BSS.

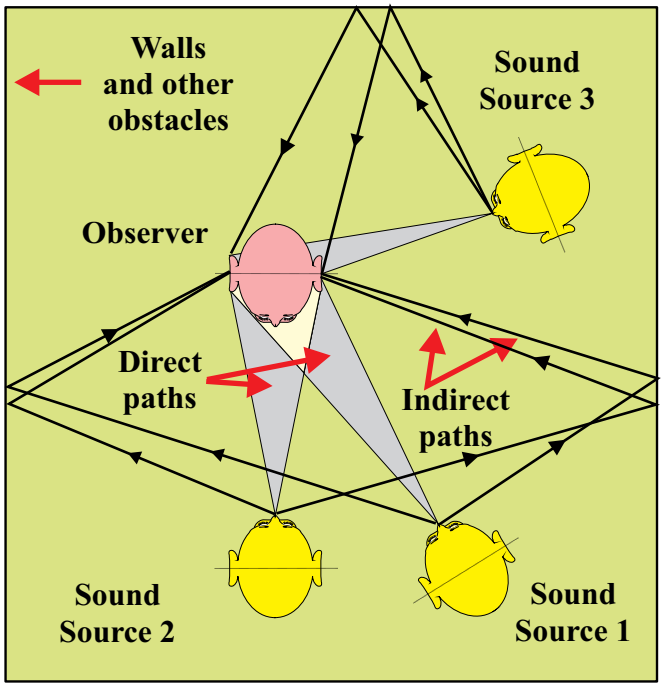
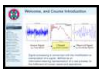


Figure 1.2: Humans turn their head in the direction of interest in order to reduce inteference from other directions; *joint detection, localisation, and enhancement*.

(a) [Huang:2008] (b) [DiBiase:2001] (c) [Wolfel:2009]

Figure 1.3: Recommended book chapters and the references therein.

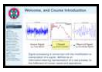
1.1.1 Structure of the Tutorial



New slide

- Recommended Texts
- Conceptual link between ASL and BSS.
- Geometry of source localisation.
- Spherical and hyperboloidal localisation.
- Estimating time-difference of arrivals (TDOAs).
- Steered beamformer response function.
- Multiple target localisation using BSS.
- Conclusions.

1.2 Recommended Texts



New slide

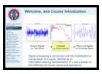
- Huang Y., J. Benesty, and J. Chen, "Time Delay Estimation and Source Localization," in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.
- Chapter 8: DiBiase J. H., H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, , Springer Berlin Heidelberg, 2001.

- Chapter 10 of Wolfel M. and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

IDENTIFIERS – *Hardback*, ISBN13: 978-0-470-51704-8

Some recent PhD thesis on the topic include:

- Zhong X., *"Bayesian framework for multiple acoustic source tracking,"* Ph.D. thesis, University of Edinburgh, 2010.
- Pertila P., *"Acoustic Source Localization in a Room Environment and at Moderate Distances,"* Ph.D. thesis, Tampere University of Technology, 2009.
- Fallon M., *"Acoustic Source Tracking using Sequential Monte Carlo,"* Ph.D. thesis, University of Cambridge, 2008.



New slide

1.3 Why Source Localisation?

A number of blind source separation (BSS) techniques rely on knowledge of the desired source position, for example:

1. Look-direction in beamforming techniques.
2. Camera steering for audio-visual BSS (including Robot Audition).
3. Parametric modelling of the mixing matrix.

Equally, a number of multi-target acoustic source localisation (ASL) techniques rely on BSS. This tutorial will look at the connections and dependencies between ASL and BSS, and discuss how they can be used together. The tutorial will cover some classical well known techniques, as well as some recent advances towards the end.

In particular, the following topics will be considered in detail:

- hyperboloidal (TDOA) based localisation methods;
- TDOA estimation methods;
- steered response power (SRP) based localisation methods;
- computationally efficient SRP methods such as stochastic region contraction (SRC);
- multi-target detection and localisation using BSS algorithms such as degenerate unmixing estimation technique (DUET);

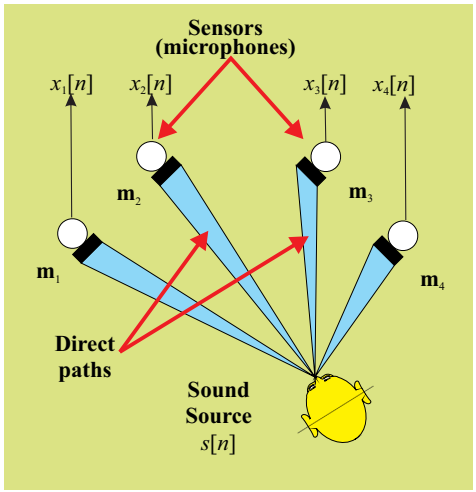
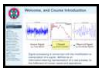


Figure 1.4: Ideal free-field model.

Figure 1.5: An ULA of microphones.

Figure 1.6: An acoustic vector sensor.



New slide

1.4 ASL Methodology

- In general, most ASL techniques rely on the fact that an impinging wavefront reaches one acoustic sensor before it reaches another.
- Most ASL algorithms are designed assuming there is no reverberation present, the *free-field assumption*; the performance of each method in the presence of reverberation will be considered after the techniques have been introduced.
- Typically, this acoustic sensor is a microphone; this tutorial will primarily consider *omni-directional pressure sensors*, and therefore many of the techniques discussed will rely on the fact there is a TDOA between the signals at different microphones.
- Other measurement types include:
  - range difference measurements;
  - interaural level difference;
  - joint TDOA and vision techniques.
- Another sensor modality might include acoustic vector sensors (AVSs) which measure both air pressure and air velocity. Useful for applications such as sniper localisation.

New slide

1.4.1 Source Localization Strategies

This section is based on

DiBiase J. H., H. F. Silverman, and M. S. Brandstein, "Robust Localization in Reverberant Rooms," in *Microphone Arrays* by M. Brandstein and D. Ward, pp. 157–180, , Springer Berlin Heidelberg, 2001.

Existing source localisation methods can loosely be divided into three generic strategies:

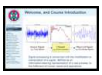
1. those based on maximising the SRP of a beamformer;
    - location estimate derivded directly from a filtered, weighted, and sum version of the signal data received at the sensors.
2. techniques adopting high-resolution spectral estimation concepts (see Stephan Weiss's talk);
    - any localisation scheme relying upon an application of the signal correlation matrix.
3. approaches employing TDOA information.
    - source locations calculated from a set of TDOA estimates measured across various combinations of microphones.

Spectral-estimation approaches See Stephan Weiss's talk :-)

TDOA-based estimators Computationally cheap, but suffers in the presence of noise and reverberation.

SBF approaches Computationally intensive, superior performance to TDOA-based methods. However, possible to dramatically reduce computational load.

1.4.2 Geometric Layout



New slide

Suppose there is a:

- sensor array consisting of N microphones located at positions $\mathbf{m}_i \in \mathbb{R}^3$, for $i \in \{0, \ldots, N-1\}$, and
- M talkers (or targets) at positions $\mathbf{x}_k \in \mathbb{R}^3$, for $k \in \{0, \ldots, M-1\}$.

The TDOA between the microphones at position $\mathbf{m}_i$ and $\mathbf{m}_j$ due to a source at $\mathbf{x}_k$ can be expressed as:

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \triangleq T_{ij}\left(\mathbf{x}_k\right) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \tag{1.1}$$

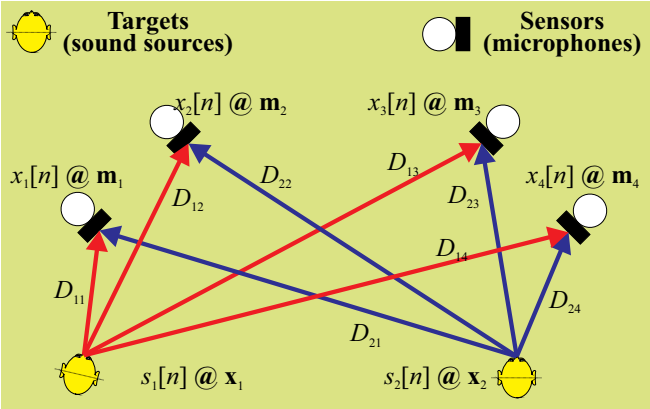


Figure 1.7: Geometry assuming a free-field model.

where c is the speed of sound, which is approximately 344 m/s. More precisely, in air, the speed of sound is given by:

$$c = 331.4 + 0.6\Theta \quad \text{m/s} \tag{1.2}$$

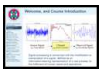
where Θ is the temperature in Centigrade or Celsius. Hence, for instance, at a temperature of 21 Celsius, then $c = 344$ m/s.

The distance from the target at $\mathbf{x}_k$ to the sensor located at $\mathbf{m}_i$ will be defined by D_{ik}, and is called the range. It is given by the expression

$$D_{ik} = |\mathbf{x}_k - \mathbf{m}_i| \tag{1.3}$$

Hence, it follows that

$$T_{ij}\left(\mathbf{x}_k\right) = \frac{1}{c}\left(D_{ik} - D_{jk}\right) \tag{1.4}$$



New slide

1.4.3 Ideal Free-field Model

- In an anechoic free-field acoustic environment, as depicted in Figure 1.4, the signal from source k, denoted by $s_k(t)$, propagates to the i-th sensor at time t according to the expression:

$$x_{ik}(t) = \alpha_{ik}\, s_k(t - \tau_{ik}) + b_{ik}(t) \tag{1.5}$$

  where $b_{ik}(t)$ denotes additive noise. Note that, in the frequency domain, this expression is given by:

$$X_{ik}\left(\omega\right) = \alpha_{ik}\, S_k\left(\omega\right)\, e^{-j\omega\,\tau_{ik}} + B_{ik}\left(\omega\right) \tag{1.6}$$

  On the assumption of **geometrical room acoustics**, which assumes high frequencies, a point sound source of single frequency ω, at position $\mathbf{x}_k$ in free space, emits a pressure wave $P_{(\mathbf{x}_k,\,\mathbf{m}_i),\,t}(\omega)$ at time t and at position $\mathbf{m}_i$:

$$P_{(\mathbf{x}_k,\,\mathbf{m}_i)}(\omega,\, t) = P_0 \frac{\exp\left[j\omega(r/c - t)\right]}{r} \tag{1.7}$$

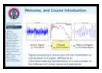
where c is the speed of sound, $t \in \mathbb{R}$ is time, and $r = |\mathbf{x}_k - \mathbf{m}_i|$, which can be seen to equate to D_{ik}.

- The additive noise source is assumed to be uncorrelated with the source signal, as well as the noise signals at the other microphones.

- The TDOA between the i-th and j-th microphone is given by:

$$\tau_{ijk} = \tau_{ik} - \tau_{jk} = T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \tag{1.8}$$

1.4.4 TDOA and Hyperboloids



New slide

It is important to be aware of the geometrical properties that arise from the TDOA relationship given in Equation 1.1:

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) = \frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c} \tag{1.9}$$

- This defines one half of a hyperboloid of two sheets, centered on the midpoint of the microphones, $\mathbf{v}_{ij} = \frac{\mathbf{m}_i + \mathbf{m}_j}{2}$. A generic diagram for the hyperboloid of twosheets is shown in Figure 1.8 and Equation 1.13. Equivalently, as shown in Sidebar 1:

$$\left(\mathbf{x}_k - \mathbf{v}_{ij}\right)^T \mathbf{V}_{ij} \left(\mathbf{x}_k - \mathbf{v}_{ij}\right) = 1 \tag{1.10}$$

where

$$\tau = cT\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right),\ \mathbf{V}_{ij} = \frac{\mathbf{I}_3 - \frac{4}{\tau^2}\boldsymbol{\mu}_{ij}\boldsymbol{\mu}_{ij}^T}{\tau^2 - |\boldsymbol{\mu}_{ij}|^2} \quad \text{and} \quad \boldsymbol{\mu}_{ij} = \frac{\mathbf{m}_i - \mathbf{m}_j}{2} \tag{1.11}$$

- For source with a large source-range to microphone-separation ratio, the hyperboloid may be well-approximated by a cone with a constant direction angle relative to the axis of symmetry. The corresponding estimated direction angle, ϕ_{ij} for the microphone pair (i, j) is given by

$$\phi_{ij} = \cos^{-1}\left(\frac{cT\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)}{|\mathbf{m}_i - \mathbf{m}_j|}\right) \tag{1.12}$$

KEYPOINT! (Hyperboloid of two sheets). General expression for a Hyperboloid of two sheets is given by:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = -1 \tag{1.13}$$

□

An example of the resulting hyperboloid for a typical case is shown in Figure 1.9, where the two-dimensional (2-D) equation is simplified in Sidebar 2. This case is for a microphone separation of $d = 0.1$, and a time-delay of $\tau_{ij} = \frac{d}{4c}$.

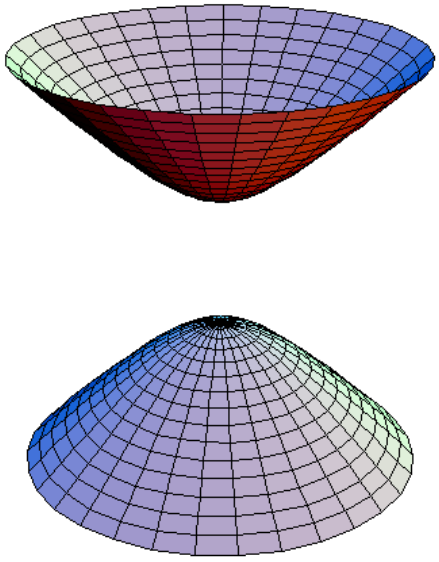


Figure 1.8: Hyperboloid of two sheets

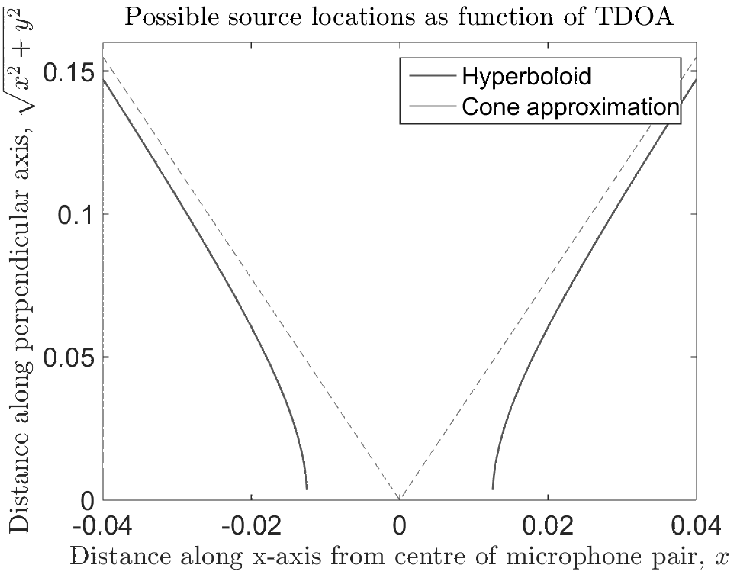


Figure 1.9: Hyperboloid, for a microphone separation of $d = 0.1$, and a time-delay of $\tau_{ij} = \frac{d}{4c}$.

Sidebar 1 Hyperboloids

Consider again Equation 1.1, but change the coordinate system to the center of the microphone pairs, such that:

$$\mathbf{x}_k = \mathbf{x} + \frac{\mathbf{m}_i + \mathbf{m}_j}{2} \tag{1.14}$$

such that:

$$\mathbf{x}_k - \mathbf{m}_i = \mathbf{x} - \underbrace{\frac{\mathbf{m}_i - \mathbf{m}_j}{2}}_{\boldsymbol{\mu}} \qquad \text{and} \qquad \mathbf{x}_k - \mathbf{m}_j = \mathbf{x} + \underbrace{\frac{\mathbf{m_i} - \mathbf{m_j}}{2}}_{\boldsymbol{\mu}} \tag{1.15}$$

The normalised-TDOA, which $\alpha = c\,\tau_{ijk}$is the actual TDOA multiplied by the speed of sound (equivalent to a range) across these two microphones can then be expressed as

$$\alpha = |\mathbf{x} - \boldsymbol{\mu}| - |\mathbf{x} + \boldsymbol{\mu}| \tag{1.16}$$

To show this is a hyperboloid, consider multiplying both sides by $|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}|$ and dividing by τ such that:

$$|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}| = \frac{1}{\alpha}\left(|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}|\right)\left(|\mathbf{x} - \boldsymbol{\mu}| - |\mathbf{x} + \boldsymbol{\mu}|\right) \tag{1.17}$$

$$= \frac{1}{\alpha}\left(|\mathbf{x} - \boldsymbol{\mu}|^2 - |\mathbf{x} + \boldsymbol{\mu}|^2\right) \tag{1.18}$$

$$|\mathbf{x} - \boldsymbol{\mu}| + |\mathbf{x} + \boldsymbol{\mu}| = -\frac{4\boldsymbol{\mu}^T\mathbf{x}}{\alpha} \tag{1.19}$$

Adding Equation 1.16 and Equation 1.19 gives:

$$2\,|\mathbf{x} - \boldsymbol{\mu}| = \alpha - \frac{4\boldsymbol{\mu}^T\mathbf{x}}{\alpha} \tag{1.20}$$

Squaring both sides again gives:

$$4\mathbf{x}^T\mathbf{x} - 8\boldsymbol{\mu}^T\mathbf{x} + 4\boldsymbol{\mu}^T\boldsymbol{\mu} = \alpha^2 - 8\boldsymbol{\mu}^T\mathbf{x} + \frac{16}{\alpha^2}\mathbf{x}^T\boldsymbol{\mu}^T\boldsymbol{\mu}^T\mathbf{x} \tag{1.21}$$

$$\mathbf{x}^T\mathbf{x} + \boldsymbol{\mu}^T\boldsymbol{\mu} = \frac{\alpha^2}{4} + \frac{4}{\alpha^2}\mathbf{x}^T\boldsymbol{\mu}\boldsymbol{\mu}^T\mathbf{x} \tag{1.22}$$

$$\mathbf{x}^T\left(\mathbf{I}_3 - \frac{4}{\alpha^2}\boldsymbol{\mu}\boldsymbol{\mu}^T\right)\mathbf{x} = \frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2 \tag{1.23}$$

finally giving:

$$\mathbf{x}^T\mathbf{V}\mathbf{x} = 1 \qquad \text{where} \qquad \mathbf{V} = \frac{\mathbf{I}_3 - \frac{4}{\alpha^2}\boldsymbol{\mu}\boldsymbol{\mu}^T}{\frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2} \tag{1.24}$$

which is the equation of an arbitrary orientated hyperboloid. The principal directions of the hyperboloid are the eigenvectors of the matrix $\mathbf{V}$. Since $\mathbf{V}$ is rank-one, it is straightforward to show that the axis of symmetry is $\boldsymbol{\mu} = \frac{\mathbf{m_i} - \mathbf{m_j}}{2}$.

Sidebar 2 Hyperboloids Example

Continuing from the derivation in Sidebar 1, suppose the microphones are at positions $\mathbf{m}_i = \begin{bmatrix}\frac{d}{2} & 0 & 0\end{bmatrix}^T$ and $\mathbf{m}_j = \begin{bmatrix}-\frac{d}{2} & 0 & 0\end{bmatrix}^T$ such that $\boldsymbol{\mu} = \begin{bmatrix}\frac{d}{2} & 0 & 0\end{bmatrix}^T$. Hence, Equation 1.24 becomes:

$$\mathbf{V} = \frac{\mathbf{I}_3 - \frac{4}{\alpha^2}\boldsymbol{\mu}\boldsymbol{\mu}^T}{\frac{\alpha^2}{4} - |\boldsymbol{\mu}|^2} \tag{1.25}$$

$$= \frac{1}{\frac{\alpha^2}{4} - \frac{d^2}{4}} \left\{ \mathbf{I}_3 - \frac{4}{\alpha^2} \begin{bmatrix} \frac{d^2}{4} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right\} \tag{1.26}$$

$$= \frac{4}{\alpha^2 - d^2} \begin{bmatrix} 1 - \frac{d^2}{\alpha^2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.27}$$

This then gives the equation of the hyperboloid as:

$$\mathbf{x}^T\mathbf{V}\mathbf{x} = 1 \tag{1.28}$$

$$\mathbf{x}^T \begin{bmatrix} 1 - \frac{d^2}{\alpha^2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} = \frac{\alpha^2 - d^2}{4} \tag{1.29}$$

$$\left(1 - \frac{d^2}{\alpha^2}\right) x^2 + y^2 + z^2 = \frac{\alpha^2 - d^2}{4} \tag{1.30}$$

$$\frac{x^2}{\left(\frac{\alpha}{2}\right)^2} - \frac{y^2 + z^2}{\frac{1}{4}\left(d^2 - \alpha^2\right)} = 1 \tag{1.31}$$

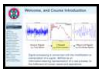
Note that the maximum TDOA will occur when the source is on the line through the two microphones, and outside of the microphones. In this case, the maximum observed delay will be $\tau_{ij} = \frac{d}{c}$ or $\alpha = d$. Hence, $d^2 - \alpha^2 \geq 0$.

Writing $r^2 = y^2 + z^2$, which are points in the $x - y$ plane on circles of radius r, this can alternatively be written as:

$$r = \frac{1}{2}\sqrt{d^2 - \alpha^2}\sqrt{\left(\frac{2x}{\alpha}\right)^2 - 1} \tag{1.32}$$

There is no solution for $x < \frac{\alpha}{2}$.

1.5 Indirect TDOA-based Methods



New slide

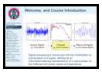
KEYPOINT! (Executive Summary). This section considers techniques which employ TDOA information directly. The section is broadly split into two sections; localising the source given TDOAs, followed by techniques for estimating TDOAs.

This is typically a two-step procedure in which:

- Typically, TDOAs are extracted using the generalised cross correlation (GCC) function, or an adaptive eigenvalue decomposition (AED) algorithm.
- A hypothesised spatial position of the target can be used to predict the expected TDOAs (or corresponding range) at the microphone.
- The error between the measured and hypothesised TDOAs is then minimised.
- Accurate and robust TDOA estimation is the key to the effectiveness of this class of ASL methods.
- An alternative way of viewing these solutions is to consider what **spatial positions** of the target could lead to the estimated TDOA.

In the following subsections, two key error functions are considered which can be optimised in a variety of methods.

1.5.1 Spherical Least Squares Error Function



New slide

KEYPOINT! (Underlying Idea). Methods using the least squares error (LSE) function relate the distance or *range* to a target, relative to each microphone, in terms of the range to a coordinate origin and the time-difference of arrival (TDOA) estimates at each microphone.

- Suppose the first microphone is located at the origin of the coordinate system, such that $\mathbf{m_0} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$.
- The range from target k to sensor i can be expressed as the range from the target to the first sensor plus a correction term:

$$D_{ik} = D_{0k} + D_{ik} - D_{0k} \tag{1.33}$$
$$= R_s + c\,T_{i0}\left(\mathbf{x}_k\right) \tag{1.34}$$

where $R_{sk} = |\mathbf{x}_k|$ is the range to the first microphone which is at the origin. This is shown in Figure 1.10.

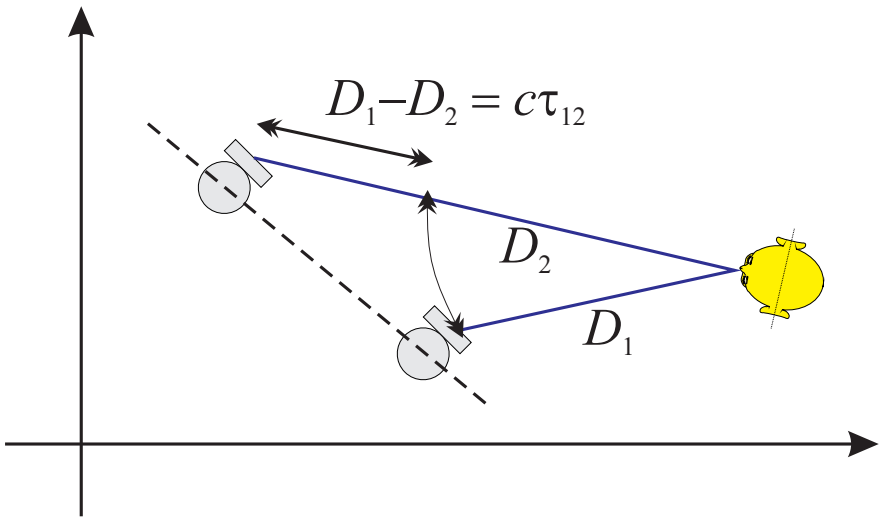


Figure 1.10: Range and TDOA relationship.

- In practice, the observations are the TDOAs and therefore, given R_{sk}, these ranges can be considered the **measurement ranges**.

  Of course, knowing R_{sk} is half the solution, but it is just one unknown at this stage. The measurements can be as

$$\hat{D}_{ik} \equiv \hat{R}_s + c\,\hat{T}_{ij} \tag{1.35}$$

- The source-sensor geometry states that the target lies on a sphere centered on the corresponding sensor. Hence,

$$\begin{aligned} D_{ik}^2 &= |\mathbf{x}_k - \mathbf{m}_i|^2 && (1.36)\\ &= \mathbf{x}_k^T\,\mathbf{x}_k - 2\mathbf{m}_i^T\,\mathbf{x}_k + \mathbf{m}_i^T\,\mathbf{m}_i && (1.37)\\ &= R_s^2 - 2\mathbf{m}_i^T\,\mathbf{x}_k + R_i^2 && (1.38) \end{aligned}$$

  where $R_i = |\mathbf{m}_i|$ is the distance of the i-th microphone to the origin.

- Define the **spherical error function** for the ith-order-microphone as the difference between the squared measured range and the squared spherical modelled range values. Using Equation 1.34 and Equation 1.38, this spherical error function can be written as:

$$\begin{aligned} \epsilon_{ik} &\triangleq \frac{1}{2}\left(\hat{D}_{ik}^2 - D_{ik}^2\right) && (1.39)\\ &= \frac{1}{2}\left\{\left(R_s + c\,\hat{T}_{i0}\right)^2 - \left(R_s^2 - 2\mathbf{m}_i^T\,\mathbf{x}_k + R_i^2\right)\right\} && (1.40)\\ &= \mathbf{m}_i^T\mathbf{x_k} + c\,R_s\,\hat{T}_{i0} + \frac{1}{2}\left(c^2\hat{T}_{i0}^2 - R_i^2\right) && (1.41) \end{aligned}$$

- Concatenating the error functions for each microphone gives the expression:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\,\mathbf{x}_k - \underbrace{(\mathbf{b}_k - R_{sk}\mathbf{d}_k)}_{\mathbf{v}_k} \tag{1.42}$$

$$\equiv \underbrace{\begin{bmatrix}\mathbf{A} & \mathbf{d}_k\end{bmatrix}}_{\mathbf{S}_k} \underbrace{\begin{bmatrix}\mathbf{x}_k \\ R_{sk}\end{bmatrix}}_{\boldsymbol{\theta}_k} - \mathbf{b}_k \tag{1.43}$$

where

$$\mathbf{A} = \begin{bmatrix}\mathbf{m}_0^T \\ \vdots \\ \mathbf{m}_{N-1}^T\end{bmatrix},\ \mathbf{d} = c\begin{bmatrix}\hat{T}_{00} \\ \vdots \\ \hat{T}_{(N-1)0}\end{bmatrix},\quad \mathbf{b}_k = \frac{1}{2}\begin{bmatrix}c^2\hat{T}_{00}^2 - R_0^2 \\ \vdots \\ c^2\hat{T}_{(N-1)0}^2 - R_{N-1}^2\end{bmatrix} \tag{1.44}$$

- The least-squares estimate (LSE) can then be obtained by forming the sum-of-squared errors term using $J = \boldsymbol{\epsilon}_i^T\boldsymbol{\epsilon}_i$ which simplifies to:

$$J(\mathbf{x}_k) = (\mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k))^T\,(\mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k)) \tag{1.45a}$$

$$J(\mathbf{x}_k,\,\boldsymbol{\theta}_k) = (\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)^T\,(\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k) \tag{1.45b}$$

- Note that as $R_{sk} = |\mathbf{x}_k|$, these parameters aren't in fact independent. Therefore, the problem to be solved can either be formulated as:
  - a nonlinear least-squares problem in $\mathbf{x}_k$ as described by Equation 1.45a;
  - a linear minimisation subject to quadratic constraints:

$$\hat{\boldsymbol{\theta}}_k = \arg\min_{\boldsymbol{\theta}_k}(\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k)^T\,(\mathbf{S}_k\boldsymbol{\theta}_k - \mathbf{b}_k) \tag{1.46}$$

subject to the constraint

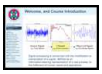
$$\boldsymbol{\theta}_k\,\Delta\,\boldsymbol{\theta}_k = 0 \qquad \text{where} \qquad \Delta = \text{diag}\,[1,\,1,\,1,\,-1] \tag{1.47}$$

The constraint $\boldsymbol{\theta}_k\,\Delta\,\boldsymbol{\theta}_k = 0$ is equivalent to

$$x_{sk}^2 + y_{sk}^2 + z_{sk}^2 = R_{sk}^2 \tag{1.48}$$

where $(x_{sk},\,y_{sk},\,z_{sk})$ are the Cartesian coordinates of the source position.

1.5.1.1 Two-step Spherical LSE Approaches



New slide

KEYPOINT! (Constrained least-squares). To avoid solving either a nonlinear or a constrained least-squares problem, it is possible to solve the problem in two steps, namely:

1. solving a LLS problem in $\mathbf{x}_k$ *assuming* the range to the target, R_{sk}, is known;
2. and then solving for R_{sk} given an estimate of $\mathbf{x}_k$ in terms of (i. t. o.) R_{sk}.

This approach is followed in the **spherical intersection (SX)** and **spherical interpolation (SI)** estimators as shown below.

- In both approaches, the range estimate is assumed known, so that the LSE can be expressed as:

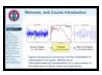
$$J\left(\mathbf{x}_k\right) = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i = \left(\mathbf{A}\,\mathbf{x}_k - \mathbf{v}_k\right)^T \left(\mathbf{A}\,\mathbf{x}_k - \mathbf{v}_k\right) \tag{1.49}$$

  Assuming an estimate of R_{sk}, denoted by $\hat{R}_{sk}$, this can be solved as

$$\hat{\mathbf{x}}_k = \mathbf{A}^\dagger\,\mathbf{v}_k = \mathbf{A}^\dagger\left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right) \quad \text{where} \quad \mathbf{A}^\dagger = \left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T \tag{1.50}$$

  Note that $\mathbf{A}^\dagger$ is the pseudo-inverse of $\mathbf{A}$.

Again, recall that the only observations are the TDOAs, $\{\hat{T}_{i0}, i \in \{0, N-1\}\}$, and that while R_{sk} is assumed known, clearly it is an unknown parameter. The differences between the following *spherical estimation* techniques essentially reduce to how the unknown range is dealt with. These are covered in the following subsections.



New slide

1.5.1.2 Spherical Intersection Estimator

This method uses the physical constraint that the range R_{sk} is the Euclidean distance to the target.

- Writing $\hat{R}_{sk}^2 = \hat{\mathbf{x}}_k^T\,\hat{\mathbf{x}}_k$, it follows that:

$$\hat{R}_{sk}^2 = \left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right)^T \mathbf{A}^{\dagger\,T}\mathbf{A}^\dagger\left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right) \tag{1.51}$$

  which can be written as the quadratic:

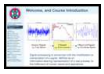
$$a\,\hat{R}_{sk}^2 + b\,\hat{R}_{sk} + c = 0 \tag{1.52}$$

  where the individual terms follow through expanding Equation 1.51. These terms are given by:

$$a = 1 - \left\|\mathbf{A}^\dagger\mathbf{d}_k\right\|^2, \quad b = 2\mathbf{b}_k\mathbf{A}^{\dagger\,T}\mathbf{A}^\dagger\mathbf{d}_k, \quad \text{and} \quad c = -\left|\mathbf{A}^\dagger\mathbf{b}_k\right\|^2 \tag{1.53}$$

- The unique, real, positive root of Equation 1.52 is taken as the SX estimator of the source range. Hence, the estimator will fail when:

  1. there is no real, positive root, or:
  2. if there are two positive real roots.



New slide

1.5.1.3 Spherical Interpolation Estimator

The SI estimator again uses the spherical LSE function, but rather than using the physically intuitive solution of *constraining* the target range relative to the origin to be the actual distance so that $R_{sk} \equiv |\mathbf{x}_k|$, it is estimated in the least-squares sense.

Consider again the **spherical error function**:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\mathbf{x}_k - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k) \tag{1.54}$$

Substituting the LSE from Equation 1.50 into this expression gives:

$$\boldsymbol{\epsilon}_{ik} = \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T\left(\mathbf{b}_k - \hat{R}_{sk}\mathbf{d}_k\right) - (\mathbf{b}_k - R_{sk}\,\mathbf{d}_k) \tag{1.55}$$

Defining the projection matrix as $\mathbf{P}_\mathbf{A} = \mathbf{I}_N - \mathbf{A}\left[\mathbf{A}^T\mathbf{A}\right]^{-1}\mathbf{A}^T$, then this may be written as:

$$\boldsymbol{\epsilon}_{ik} = R_{sk}\,\mathbf{P}_\mathbf{A}\mathbf{d}_k - \mathbf{P}_\mathbf{A}\mathbf{b}_k \tag{1.56}$$

Minimising the LSE using the normal equations gives:

$$R_{sk} = \left(\mathbf{d}_k^T\,\mathbf{P}_\mathbf{A}^T\,\mathbf{P}_\mathbf{A}\mathbf{d}_k\right)^{-1}\,\mathbf{d}_k^T\mathbf{P}_\mathbf{A}^T\mathbf{P}_\mathbf{A}\mathbf{b}_k \tag{1.57}$$

However, the **projection matrix** is symmetric and idempotent, such that $\mathbf{P}_\mathbf{A} = \mathbf{P}_\mathbf{A}^T$ and $\mathbf{P}_\mathbf{A}\mathbf{P}_\mathbf{A} = \mathbf{P}_\mathbf{A}$. This means that the sum-of-squared errors simplifies to:

$$R_{sk} = \left(\mathbf{d}_k^T\,\mathbf{P}_\mathbf{A}\mathbf{d}_k\right)^{-1}\,\mathbf{d}_k^T\mathbf{P}_\mathbf{A}\mathbf{b}_k \tag{1.58}$$

or alternatively, since the quantity in the inverse is a scaler,

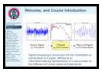
$$R_{sk} = \frac{\mathbf{d}_k^T\mathbf{P}_\mathbf{A}\mathbf{b}_k}{\mathbf{d}_k^T\,\mathbf{P}_\mathbf{A}\mathbf{d}_k} \tag{1.59}$$

Substituting back into the LSE for the target position given in Equation 1.50 gives the final estimator:

$$\hat{\mathbf{x}}_k = \mathbf{A}^\dagger\left(\mathbf{I}_N - \mathbf{d}_k\frac{\mathbf{d}_k^T\mathbf{P}_\mathbf{A}}{\mathbf{d}_k^T\,\mathbf{P}_\mathbf{A}\mathbf{d}_k}\right)\mathbf{b}_k \tag{1.60}$$

This approach is said to perform better, but is computationally slightly more complex than the SX estimator.

1.5.1.4 Other Approaches

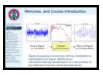


New slide

There are several other approaches to minimising the spherical LSE function defined in Equation 1.45.

- In particular, the **linear-correction** LSE solves the constrained minimization problem using Lagrange multipliers in a two stage process.
- For further information, see: Huang Y., J. Benesty, and J. Chen, "Time Delay Estimation and Source Localization," in *Springer Handbook of Speech Processing* by J. Benesty, M. M. Sondhi, and Y. Huang, pp. 1043–1063, , Springer, 2008.

1.5.2 Hyperbolic Least Squares Error Function



New slide

KEYPOINT! (Underlying Concept). Suppose that for each pair of microphones i and j, a TDOA corresponding to source k is somehow estimated, and this is denoted by τ_{ijk}. One approach to ASL is to minimise the total error between the measured TDOAs and the TDOAs predicted by the geometry *given* an assumed target position.

- If a TDOA is estimated between two microphones i and j, then the error between this and modelled TDOA is given by Equation 1.1:

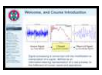
$$\epsilon_{ij}(\mathbf{x}_k) = \tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \tag{1.61}$$

  where the error is considered as a function of the source position $\mathbf{x}_k$.

- The total error as a function of target position

$$J(\mathbf{x}_k) = \sum_{i=1}^{N} \sum_{j \neq i=1}^{N} \left(\tau_{ijk} - T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)\right)^2 \tag{1.62}$$

- Unfortunately, since $T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right)$ is a nonlinear function of $\mathbf{x}_k$, the minimum LSE does not possess a closed-form solution.



New slide

1.5.2.1 Linear Intersection Method

KEYPOINT! (Underlying Concept). The linear intersection (LI) algorithm works by utilising a *sensor quadruple* with a common midpoint, which allows a bearing line to be deduced from the intersection of two cones which approximate the hyperboloid. The spatial position that minimises the distance between these bearing lines a the point of nearest intersection is considered the target position.

- Given the bearing lines, it is possible to calculate the points $\mathbf{s}_{ij}$ and $\mathbf{s}_{ji}$ on two bearing lines which give the closest intersectionas illustrated in Figure 1.12. This is basic gemoentry, and for a detailed analysis, see [Brandstein:1997].

- The trick is to note that given these points $\mathbf{s}_{ij}$ and $\mathbf{s}_{ji}$, the theoretical TDOA, $T\left(\mathbf{m}_{1i},\, \mathbf{m}_{2i},\, \mathbf{s}_{ij}\right)$, can be compared with the observed TDOA.

  This will then lead to a weighted location estimate, where the weights are related to the likelihood of the target position given the observed TDOA.

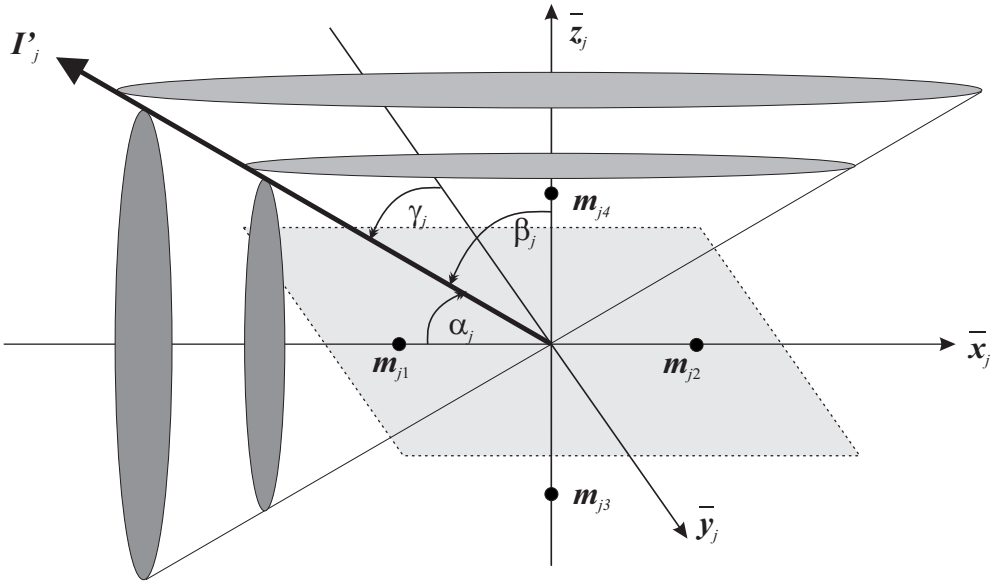


Figure 1.11: Quadruple sensor arrangement and local Cartesian coordinate system.

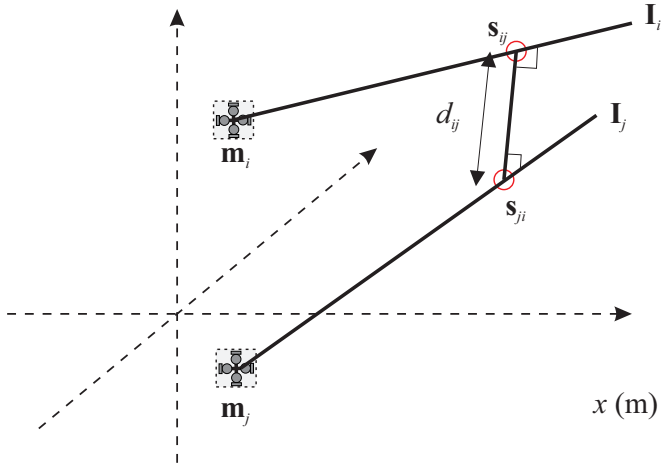
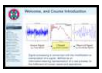


Figure 1.12: Calculating the points of closest intersection.



1.5.3 TDOA estimation methods

New slide

Two key methods for TDOA estimation are using the GCC function and the AED algorithm.

GCC algorithm most popular approach assuming an ideal free-field movel. It has the advantages that

- computationally efficient, and hence short decision delays;
- perform fairly well in moderately noisy and reverberant environments.

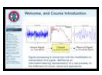
However, GCC-based methods

- fail when room reverberation is high;
- focus of current research is on combating the effect of room reverberation.

AED Algorithm Approaches the TDOA estimation approach from a different point of view from the *traditional* GCC method.

- adopts a reverberant rather than free-field model;
- computationally more expensive than GCC;
- can fail when there are common-zeros in the room impulse response (RIR).

Note that both methods assume that the signals received at the microphones arise as the result of a single source, and that if there are multiple sources, the signals will first need to be separated into different contributions of the individual sources.



1.5.3.1 GCC TDOA estimation

New slide

The GCC algorithm proposed by *Knapp and Carter* is the most widely used approach to TDOA estimation.

- The TDOA estimate between two microphones i and j is obtained as the time lag that maximises the cross-correlation between the filtered versions of the microphone outputs:

$$\hat{\tau_{ij}} = \arg\max_{\ell} r_{x_i x_j}[\ell] \tag{1.63}$$

where the signal received at microphone i is given by $x_i[n]$, and where x_i should not be confused with the location of the source k, which is denoted by $\mathbf{x}_k = [x_k,\, y_k,\, z_k]^T$.

- The cross-correlation function is given by

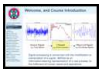
$$r_{x_i\,x_j}[\ell] = \mathcal{F}^{-1}\left(\Psi_{x_1x_2}\left(e^{j\omega T_s}\right)\right) \tag{1.64}$$
$$= \mathcal{F}^{-1}\left(\Phi\left(e^{j\omega T_s}\right) P_{x_1x_2}\left(e^{j\omega T_s}\right)\right) \tag{1.65}$$
$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Psi_{x_1x_2}\left(e^{j\omega T_s}\right)\, e^{j\ell\omega T}\, d\omega \tag{1.66}$$
$$= \int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}} \Phi\left(e^{j\omega T_s}\right) P_{x_1x_2}\left(e^{j\omega T_s}\right)\, e^{j\ell\omega T}\, d\omega \tag{1.67}$$

  where the cross-power spectral density (CPSD) is given by

$$P_{x_1x_2}\left(e^{j\omega T_s}\right) = \mathbb{E}\left[X_1\left(e^{j\omega T_s}\right) X_2\left(e^{j\omega T_s}\right)\right] \tag{1.68}$$

  The CPSD can be estimated in a variety of means. The choice of the filtering term or frequency domain weighting function, $\Phi\left(e^{j\omega T_s}\right)$, leads to a variety of different GCC methods for TDOA estimation. In Section 1.5.3.3 , some of the popular approaches are listed, but only one is covered in detail, namely the phase transform (PHAT).

1.5.3.2 CPSD for Free-Field Model



New slide

For the free-field model in Equation 1.5 and Equation 1.6, it follows that for $i \neq j$ the CPSD in Equation 1.68 is given by:

$$P_{x_ix_j}(\omega) = \mathbb{E}\left[X_j(\omega)\, X_j(\omega)\right] \tag{1.69}$$
$$= \mathbb{E}\left[\left(\alpha_{ik}\, S_k(\omega)\, e^{-j\omega\,\tau_{ik}} + B_{ik}(\omega)\right)\left(\alpha_{jk}\, S_k(\omega)\, e^{-j\omega\,\tau_{kk}} + B_{jk}(\omega)\right)\right] \tag{1.70}$$
$$= \alpha_{ik}\alpha_{jk} e^{-j\omega\, T(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k)} \mathbb{E}\left[\left|S_k(\omega)\right|^2\right] \tag{1.71}$$

where $\mathbb{E}\left[B_{ik}(\omega)\, B_{jk}(\omega)\right] = 0$ and $\mathbb{E}\left[B_{ik}(\omega)\, S_k(\omega)\right] = 0$ due to the noise being uncorrelated with the source signal and noise signals.

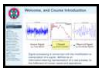
- In particular, note that it follows:

$$\angle P_{x_ix_j}(\omega) = -j\omega\, T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \mathbf{x}_k\right) \tag{1.72}$$

  In otherwords, all the TDOA information is conveyed in the phrase rather than the amplitude of the CPSD. This therefore suggests that the weighting function can be chosen to remove the amplitude information.

  These equations can be converted to discrete time as appropriate.

1.5.3.3 GCC Processors



New slide

The most common choices for the GCC weighting term are listed in the table below. In particular, the PHAT is considered in detail.

Processor Name	Frequency Function
Cross Correlation	1
PHAT	$\frac{1}{\lvert P_{x_1x_2}\left(e^{j\omega T_s}\right)\rvert}$
Roth Impulse Response	$\frac{1}{P_{x_1x_1}\left(e^{j\omega T_s}\right)}$ or $\frac{1}{P_{x_2x_2}\left(e^{j\omega T_s}\right)}$
SCOT	$\frac{1}{\sqrt{P_{x_1x_1}\left(e^{j\omega T_s}\right)P_{x_2x_2}\left(e^{j\omega T_s}\right)}}$
Eckart	$\frac{P_{s_1s_1}\left(e^{j\omega T_s}\right)}{P_{n_1n_1}\left(e^{j\omega T_s}\right)P_{n_2n_2}\left(e^{j\omega T_s}\right)}$
Hannon-Thomson or ML	$\frac{\left\lvert\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\right\rvert^2}{\lvert P_{x_1x_2}\left(e^{j\omega T_s}\right)\rvert\left(1-\lvert\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)\rvert^2\right)}$

where $\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)$ is the normalised CPSD or **coherence function** is given by

$$\gamma_{x_1x_2}\left(e^{j\omega T_s}\right)=\frac{P_{x_1x_2}\left(e^{j\omega T_s}\right)}{\sqrt{P_{x_1x_1}\left(e^{j\omega T_s}\right)P_{x_2x_2}\left(e^{j\omega T_s}\right)}} \tag{1.73}$$

The PHAT-GCC approach can be written as:

$$r_{x_i\,x_j}[\ell]=\int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}}\Phi\left(e^{j\omega T_s}\right)P_{x_1x_2}\left(e^{j\omega T_s}\right)\,e^{j\ell\omega T}\,d\omega \tag{1.74}$$

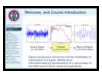
$$=\int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}}\frac{1}{\lvert P_{x_1x_2}\left(e^{j\omega T_s}\right)\rvert}\lvert P_{x_1x_2}\left(e^{j\omega T_s}\right)\rvert e^{j\angle P_{x_1x_2}\left(e^{j\omega T_s}\right)}\,e^{j\ell\omega T}\,d\omega \tag{1.75}$$

$$=\int_{-\frac{\pi}{T_s}}^{\frac{\pi}{T_s}}e^{j\left(\ell\omega T+\angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right)}\,d\omega \tag{1.76}$$

$$=\delta\left(\ell\,T_s+\angle P_{x_1x_2}\left(e^{j\omega T_s}\right)\right) \tag{1.77}$$

$$=\delta(\ell\,T_s-T\left(\mathbf{m}_i,\,\mathbf{m}_j,\,\mathbf{x}_k\right)) \tag{1.78}$$

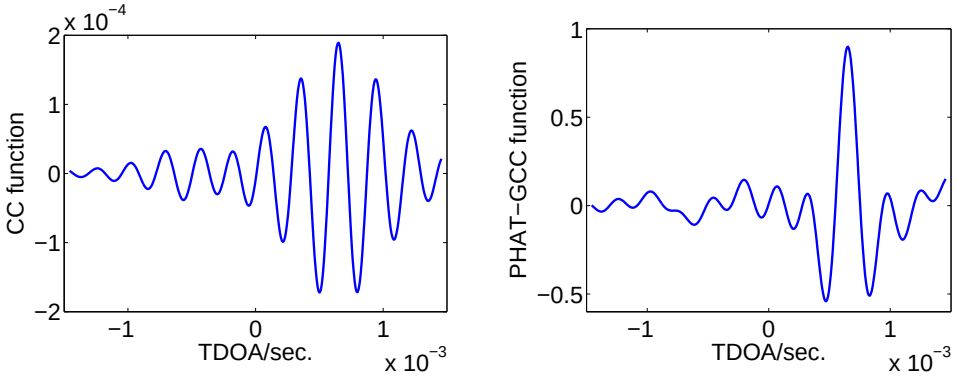
- In the absence of reverberation, the GCC-PHAT (GCC-PHAT) algorithm gives an impulse at a lag given by the TDOA divided by the sampling period.



New slide

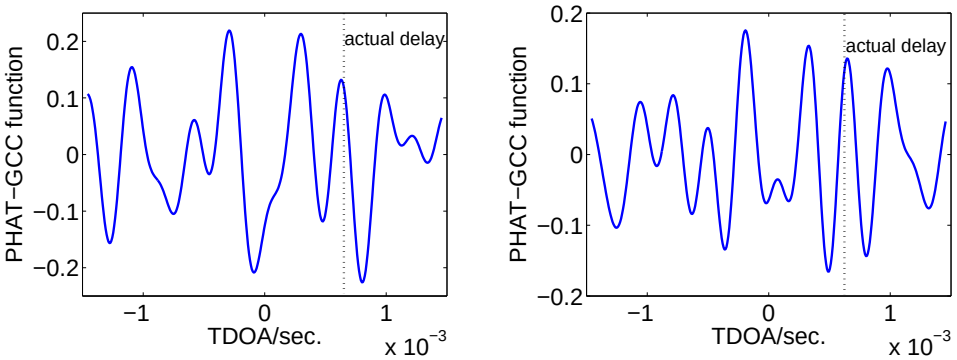
1.5.3.4 Adaptive Eigenvalue Decomposition

KEYPOINT! (Underlying Concept). The AED algorithm adopts the real reverberant rather than free-field model. The AED algorithm actually amounts to a **blind channel identification** problem, which then seeks to identify the channel coefficients corresponding to the direct path elements.



(a) Cross-correlation function. (b) GCC-PHAT function

Figure 1.13: Normal cross-correlation and GCC-PHAT functions for a frame of speech.



(a) GCC-PHAT in a reverberant environment, $\rho = 08$. The ground truth of TDOA is 0.64 ms.

(b) GCC-PHAT in a noisy environment, SNR $= 0$ dB.

Figure 1.14: The effect of reverberation and noise on the GCC-PHAT can lead to poor TDOA estimates.

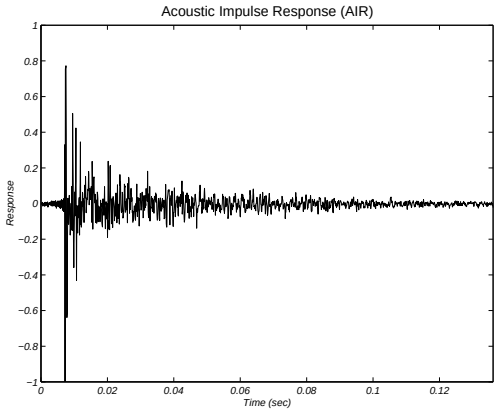


Figure 1.15: A typical room acoustic impulse response.

- Suppose that the acoustic impulse response (AIR) between source k and i is given by $h_{ik}[n]$ such that

$$x_{ik}[n] = \sum_{m=-\infty}^{\infty} h_{ik}[n-m]\, s_k[m] + b_{ik}[n] \tag{1.79}$$

  then the TDOA between microphones i and j is:

$$\tau_{ijk} = \left\{ \arg\max_{\ell} |h_{ik}[\ell]| \right\} - \left\{ \arg\max_{\ell} |h_{jk}[\ell]| \right\} \tag{1.80}$$

  This assumes a minimum-phase system, but can easily be made robust to a non-minimum-phase system.

- Reverberation plays a major role in ASL and BSS.
- Consider reverberation as the sum total of all sound reflections arriving at a certain point in a room after room has been excited by impulse.

Trivia: Perceive early reflections to reinforce direct sound, and can help with speech intelligibility. It can be easier to hold a conversation in a closed room than outdoors

- Room transfer functions are often nonminimum-phase since there is more energy in the reverberant component of the RIR than in the component corresponding to sound travelling along a direct path.

- Therefore AED will need to consider multiple peaks in the estimated AIR.

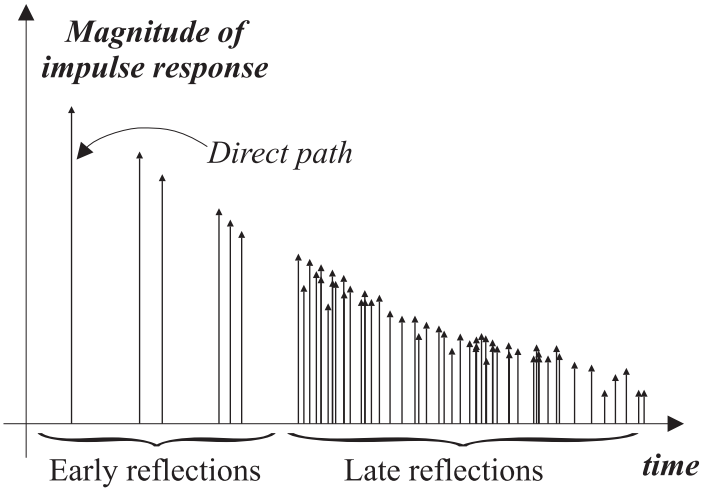


Figure 1.16: Early and late reflections in an AIR.

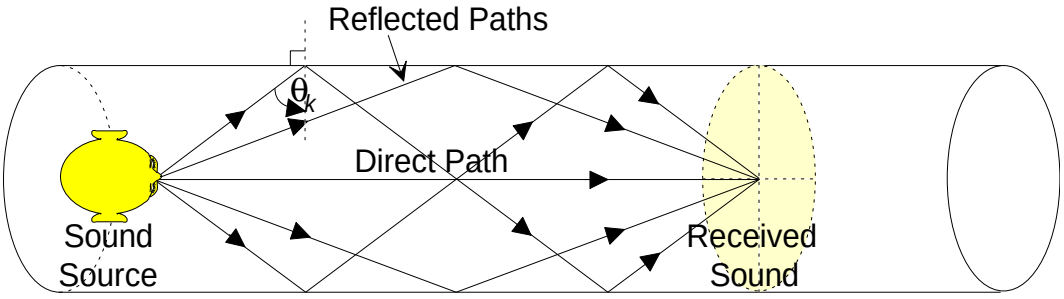
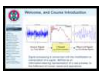


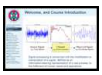
Figure 1.17: In an infinitely long cylindrical tube, the reverberant energy is greater than the energy contained in the sound travelling along a direct path, thus demonstrating the nonminimum-phase properties of room acoustics.



1.6 Direct Localisation Methods

New slide

- Direct localisation methods have the advantage that the relationship between the measurement and thestate is linear.
- However, extracting the position measurement requires a multi-dimensional search over the state space and is usually computationally expensive.



1.6.1 Steered Response Power Function

New slide

KEYPOINT! (Underlying Concept). The steered beamformer (SBF) or SRP function is a measure of correlation across *all pairs* of microphone signals for a set of relative delays that arise from a hypothesised source location.

The frequency domain **delay-and-sum beamformer** steered to a spatial position $\hat{\mathbf{x}}_k$ such that $\hat{\tau}_{pk} = |\hat{\mathbf{x}} - \mathbf{m}_p|$, using the notation in Equation 1.8, is given by:

$$S\left(\hat{\mathbf{x}}\right) = \int_{\Omega} \left| \sum_{p=1}^{N} W_p\left(e^{j\omega T_s}\right)\, X_p\left(e^{j\omega T_s}\right)\, e^{j\omega\,\hat{\tau}_{pk}} \right|^2 d\omega \tag{1.81}$$

Expanding and rearranging the order of integration and summation gives:

$$S\left(\hat{\mathbf{x}}\right) = \int_{\Omega} \sum_{p=1}^{N}\sum_{q=1}^{N} W_p\left(e^{j\omega T_s}\right)\, W_q^*\left(e^{j\omega T_s}\right) X_p\left(e^{j\omega T_s}\right) X_q^*\left(e^{j\omega T_s}\right)\, e^{j\omega\left(\hat{\tau}_{pk}-\hat{\tau}_{qk}\right)}\, d\omega \tag{1.82}$$

Taking expectations of both sides and setting $\Phi_{pq}\left(e^{j\omega T_s}\right) = W_p\left(e^{j\omega T_s}\right)\, W_q^*\left(e^{j\omega T_s}\right)$ gives

$$\mathbb{E}\left[S\left(\hat{\mathbf{x}}\right)\right] = \sum_{p=1}^{N}\sum_{q=1}^{N} \int_{\Omega} \Phi_{pq}\left(e^{j\omega T_s}\right)\, P_{x_p x_q}\left(e^{j\omega T_s}\right)\, e^{j\omega\hat{\tau}_{pqk}}\, d\omega \tag{1.83}$$

$$= \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}[\hat{\tau}_{pqk}] \equiv \sum_{p=1}^{N}\sum_{q=1}^{N} r_{x_i x_j}\left[\frac{|\mathbf{x}_k - \mathbf{m}_i| - |\mathbf{x}_k - \mathbf{m}_j|}{c}\right] \tag{1.84}$$

In other words, the SRP is the sum of all possible pairwise GCC functions evaluated at the time delays hypothesised by the target position. This is discussed in Section 1.6.2.

1.6.2 Conceptual Intepretation of SRP

New slide

Equation 1.84 gives an elegant conceptual intepretation of the SBF function. Given a candidate spatial position $\hat{\mathbf{x}}_k$, the corresponding TDOA at microphones i and j can be calculated using Equation 1.9:

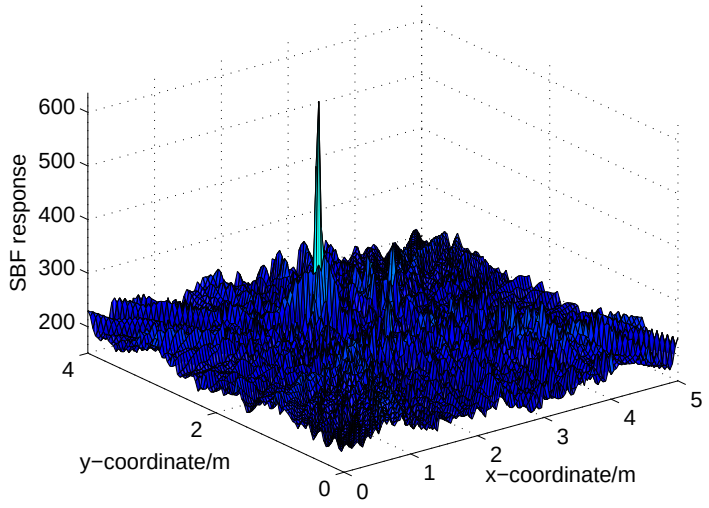


Figure 1.18: SBF response from a frame of speech signal. The integration frequency range is 300 to 3500 Hz (see Equation 1.84). The true source position is at $[2.0, 2.5]m$. The grid density is set to 40 mm.

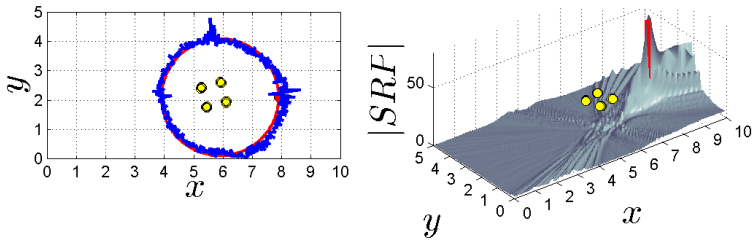


Figure 1.19: An example video showing the SBF changing as the source location moves.

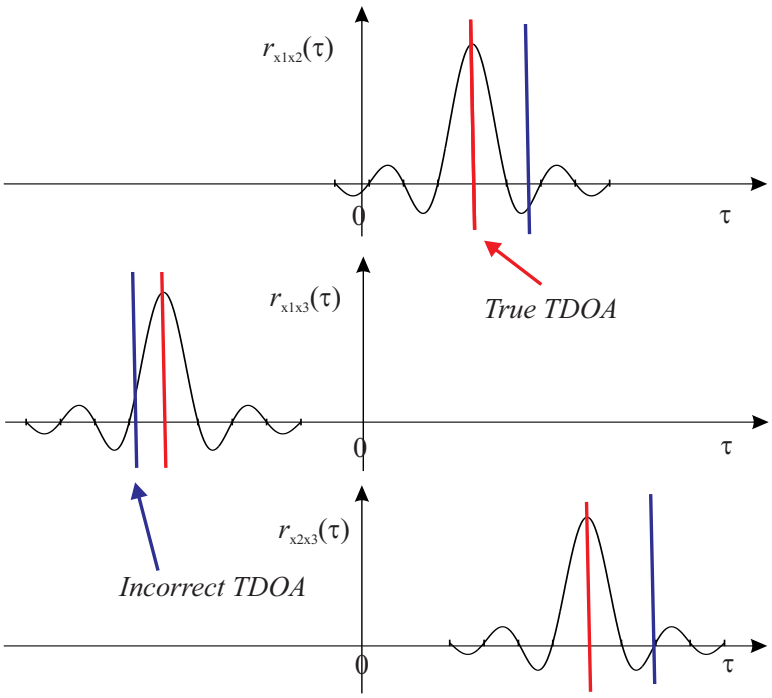


Figure 1.20: GCC-PHAT for different microphone pairs.

$$T\left(\mathbf{m}_i,\, \mathbf{m}_j,\, \hat{\mathbf{x}}_k\right) = \frac{|\hat{\mathbf{x}}_k - \mathbf{m}_i| - |\hat{\mathbf{x}}_k - \mathbf{m}_j|}{c} \tag{1.85}$$

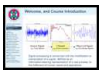
Since the SBF function in Equation 1.84 is a linear combination of the GCC-PHAT functions, then if $\hat{\mathbf{x}}_k$ is correct, then the GCC-PHAT functions should return a large peak. If $\hat{\mathbf{x}}_k$ is incorrect, then the GCC-PHAT functions return smaller values, and therefore the SBF function in Equation 1.84 is smaller.

2

Blind Source Localisation

This handout considers multi-target localisation using blind source separation (BSS) techniques.

2.1 DUET Algorithm



New slide

KEYPOINT! (Summary). The degenerate unmixing estimation technique (DUET) algorithm is an approach to blind source separation (BSS) that ties in neatly to acoustic source localisation (ASL). Under certain assumptions and circumstances, it is possible to separate more than two sources using only two microphones.

- DUET is based on the assumption that for a set of signals $x_k[t]$, their time-frequency representations (TFRs) are predominately non-overlapping. This condition is referred to as W-disjoint orthogonality (WDO), and can be stated as follows:

$$S_p(\omega,\, t)\; S_q(\omega,\, t) = 0 \,\forall p \neq q,\, \forall t,\, \omega \tag{2.1}$$

The WDO property is clearly shown in Figure 2.1, where the spectrograms of *clean* speech mixtures are sparse and disjoint. For two speech signals, the product of the corresponding spectrograms is zero at the most area on the time-frequency (TF) domain.

Consider taking then, a particular TF-bin, $(\omega,\, t)$, where source p is known to be active. The two received signals at microphones i and j in *that TF-bin* can be written in the

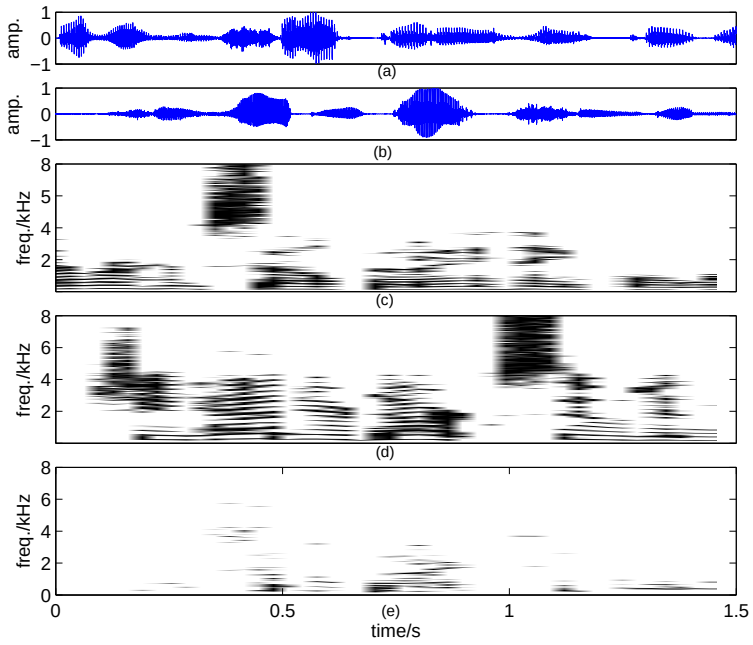
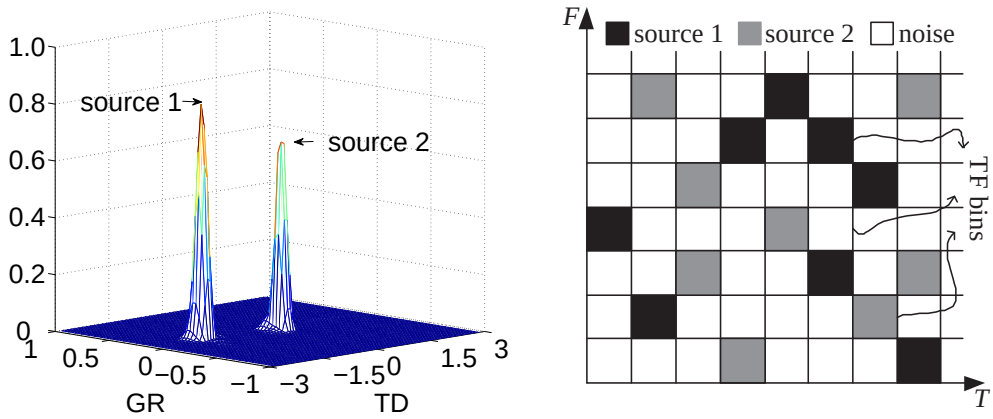


Figure 2.1: W-disjoint orthogonality of two speech signals. Original speech signal (a) $s_1[t]$ and (b) $s_2[t]$; corresponding STFTs (c) $|S_1(\omega, t)|$ and (d) $|S_2(\omega, t)|$; (e) product of the two spectrogram $|S_1(\omega, t)\, S_2(\omega, t)|$.



(a) Histogram of two sources in an anechoic environment.

(b) TF-mask for each source.

Figure 2.2: Illustration of the underlying idea in DUET.

TF-domain as:

$$\begin{aligned} X_{ip}(\omega,\, t) &= \alpha_{ip}\, e^{-j\omega\,\tau_{ip}}\, S_p(\omega,\, t) + B_i(\omega,\, t) \\ X_{jp}(\omega,\, t) &= \alpha_{jp}\, e^{-j\omega\,\tau_{jp}}\, S_p(\omega,\, t) + B_j(\omega,\, t) \end{aligned} \tag{2.2}$$

Taking the ratio of these expressions and ignoring the noise terms gives:

$$H_{ikp}(\omega,\, t) \triangleq \frac{X_{ip}(\omega,\, t)}{X_{jp}(\omega,\, t)} = \frac{\alpha_{ip}}{\alpha_{jp}}\, e^{-j\omega\tau_{ijp}} \tag{2.3}$$

where, again, τ_{ijp} is the time-difference of arrival (TDOA) of the signal contribution due to source p between microphones i and j.

KEYPOINT! (Which TF-bins belong to which source?). Of course, which TF-bins belong to which source is unknown, as the source signal and spectrum is unknown. However, if the magnitude and phase terms of the ratio in Equation 2.3 are *histogrammed* over all TF-bins, peaks will occur a distinct magnitude-phase positions, each peak corresponding to a different source.

Hence,

$$\tau_{ijp} = -\frac{1}{\omega} \arg H_{ikp}(\omega,\, t)\,, \qquad \text{and} \qquad \frac{\alpha_{ip}}{\alpha_{jp}} = |H_{ikp}(\omega,\, t)| \tag{2.4}$$

This leads to the essentials of the DUET method which are:

1. Construct the TF representation of both mixtures.
2. Take the ratio of the two mixtures and extract local mixing parameter estimates.
3. Combine the set of local mixing parameter estimates into N pairings corresponding to the true mixing parameter pairings.

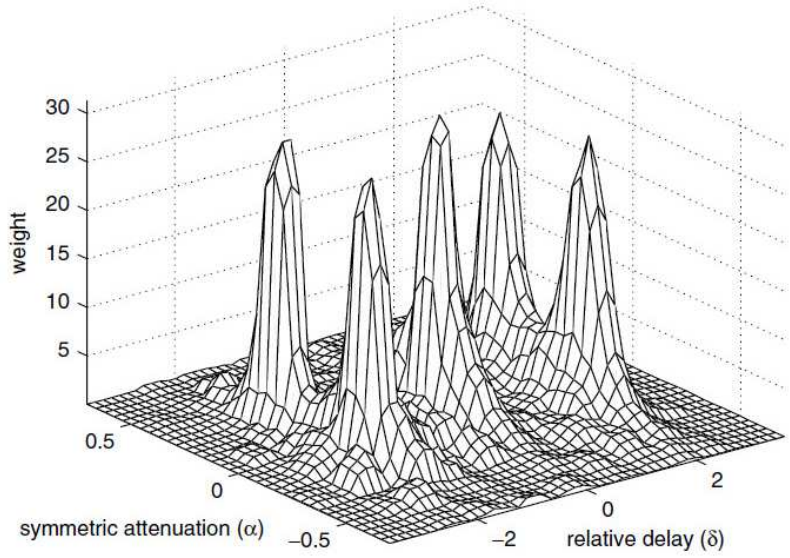
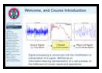


Figure 2.3: DUET for multiple sources.

4. Generate one binary mask for each determined mixing parameter pair corresponding to the TF-bins which yield that particular mixing parameter pair.

5. Demix the sources by multiplying each mask with one of the mixtures.

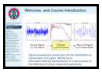
6. Return each demixed TFR to the time domain.



2.1.1 Effect of Reverberation and Noise

New slide

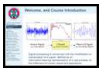
A number of papers have analysed the validity of the WDO property, and anechoic speech often satisfies this. However, while the TFR of speech is very clear in this case, the TFR becomes smeared due to revebereration and noise.



2.1.2 Estimating multiple targets

New slide

The underlying idea is shown in Figure 2.5 and Figure 2.6.

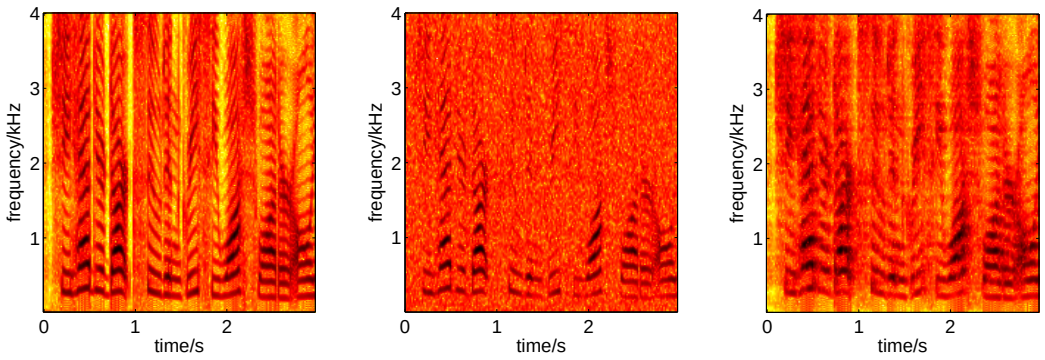


2.2 Further Topics

New slide

- Reduction in complexity of calculating steered response power (SRP). This includes stochastic region contraction (SRC) and hierarchical searches.

- Multiple-target tracking (see Daniel Clark's Notes)



(a) An anechoic environment. (b) A reverberant environment. (c) A noisey environment.

Figure 2.4: The TFR is very clear in the anechoic environment but smeared around by the reverberation and noise.

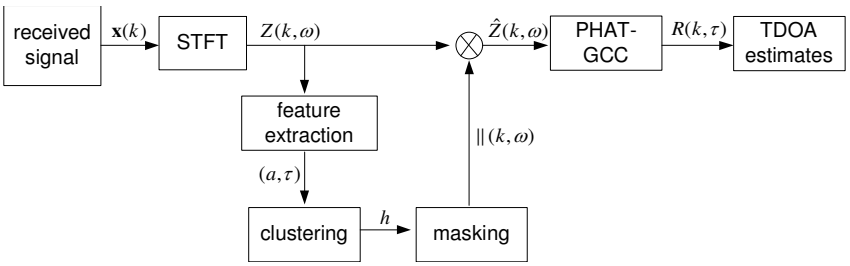


Figure 2.5: Flow diagram of the DUET-GCC approach. Basically, the speech mixtures are separated by using the DUET in the TF domain, and the PHAT-GCC is then employed for the spectrogram of each source to estimate the TDOAs.

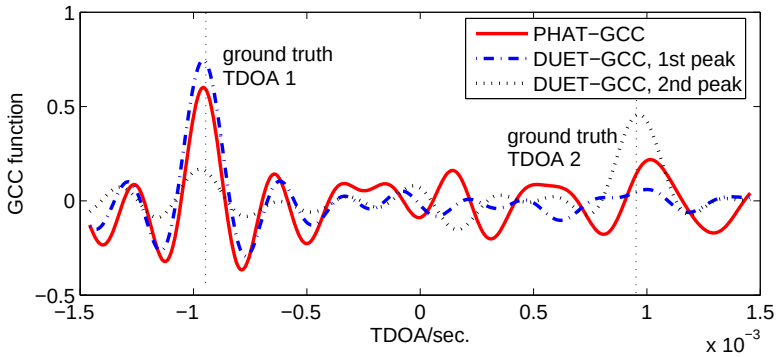
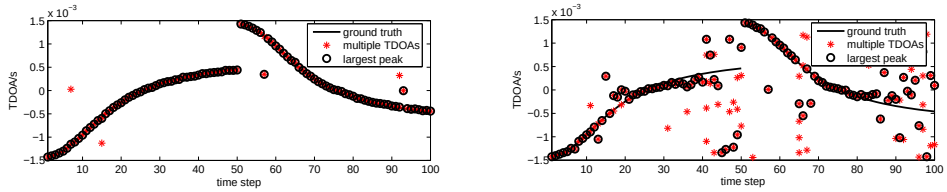


Figure 2.6: GCC function from DUET approach and traditional PHAT weighting. Two sources are located at (1.4, 1.2)m and (1.4, 2.8)m respectively. The GCC function is estimated from the first microphone pair (microphone 1 and microphone 2). The ground truth TDOAs are 0.95 ms.



(a) (b)

Figure 2.7: Acoustic source tracking and localisation.

- Simultaneous (self-)localisation and tracking; estimating sensor and target positions from a moving source.

- Joint ASL and BSS.
- Explicit signal and channel modelling! (None of the material so forth cares whether the signal is speech or music!)
- Application areas such as gunshot localisation; other sensor modalities; diarisation.