Speaker location and acoustic event detection given a distributed microphone network

Maurizio Omologo
ITC-irst, Trento, Italy

* Some of the activities and results described in this talk represent achievements obtained at ITC-irst under CHIL (Computers In the Human Interaction Loop) Integrated Project (IP 506909) funded by the European Commission's Sixth Framework Programme
(for more details on CHIL, please consult *http://chil.server.de*
or contact Alex Waibel and Rainer Stiefelhagen at Karlsruhe University, Germany).

Outline

- **Part I: Introduction**
  - The CHIL project: general objectives
  - Foreseen applicative contexts
  - Acoustic scene analysis: distributed microphone networks
- **Part II: Speaker Location**
  - Common approaches and basic techniques
  - Global Coherence Field (GCF) and Oriented GCF (OGCF)
  - Experimental results and NIST benchmarking
  - Demo
- **Part III: Other techniques for Acoustic Scene Understanding**
  - Speech/Noise Source Activity Detection
  - Multi-microphone based F0 Estimation + Demo
  - Acoustic Event Classification + Demo
  - Distant-talking Speech Recognition + Demo
  - ASR and Understanding given a Microphone Network
- **Part IV: Future work**

Outline

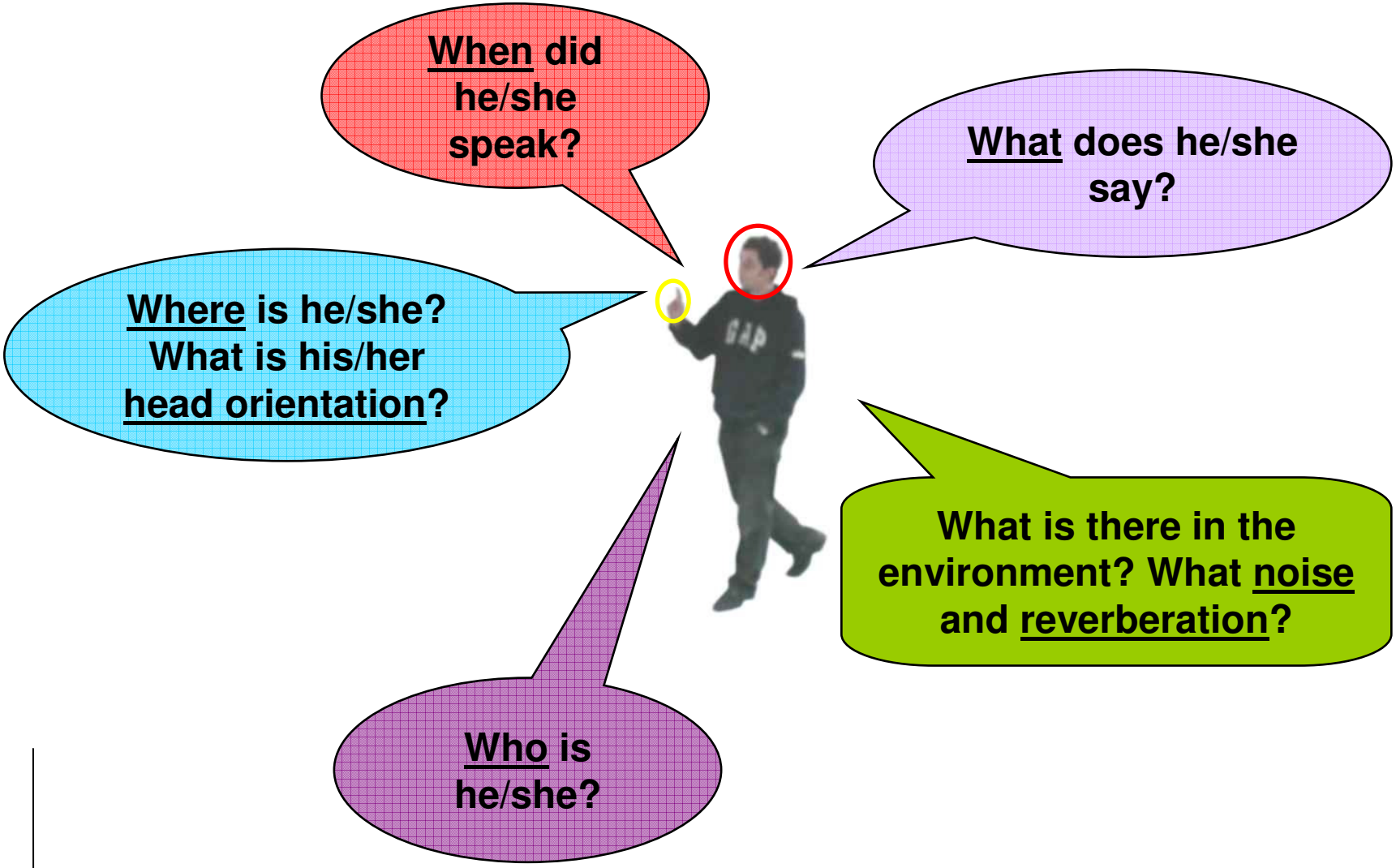
- **Part I: Introduction**
  - The CHIL project: general objectives
  - Foreseen applicative contexts
  - Acoustic scene analysis: distributed microphone networks

Some of the objectives of the CHIL Project

- Multi-sensor, multi-modal processing for robust person location, tracking and identification under unconstrained conditions (acoustic noise, visual occlusion, non-frontality, illumination variation)
- Audio-visual far-field source separation, speech recognition, scene analysis for activity detection and description
- Body expression at various scales (body movements, gestures and postures)

- Three main groups of technologies:
  1. **Who and Where**
  2. **What**
  3. **Why and How**
- Experimental activities by using same training/test data, collected across the CHIL rooms available at various partner sites.

CHIL Project: the acoustic scene analysis task



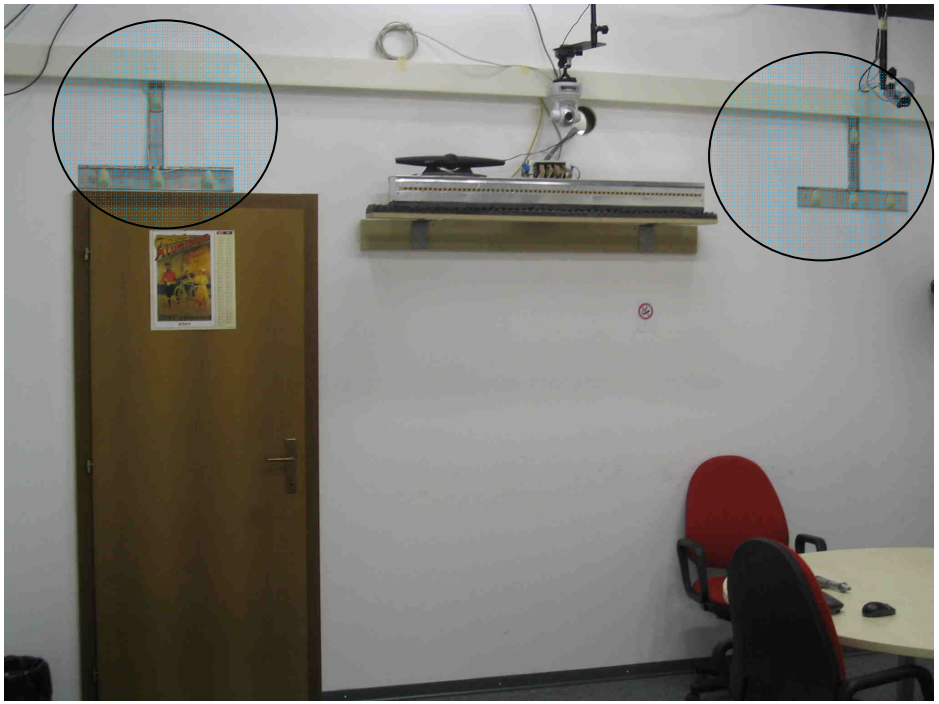
Foreseen Applicative Contexts

- Smart-rooms (lectures, meetings, surgery rooms, etc)
- Smart home (e.g., television and other appliance control)
- Camera-tracking for video-conferencing
- Surveillance
- Security
- Robotics
- Distant-talking speech recognition for:
  - banking,
  - elderly and disabled assistance,
  - manufacturing,
  - process control,
  - dictation/data entry for document creation, etc

Distributed Microphone Networks in CHIL

Distributing microphone arrays in space allows to cover any area in a more effective way. Some issues:

- Type of microphone
- Number of sensors
- Array geometries
- Synchronous sampling and processing
- Integration with video sensors



CHIL room at ITC-irst:

- 4.75 m x 5.93 m x ~4.5 m size
- Reverberation time ~ 700 ms
- Microphone Network consists of 7 reverse T-shaped arrays, 1 NIST MarkIII/ IRST array, 1 commercial array, 4 close-talking and 8 table microphones

CHIL room at Karlsruhe University:

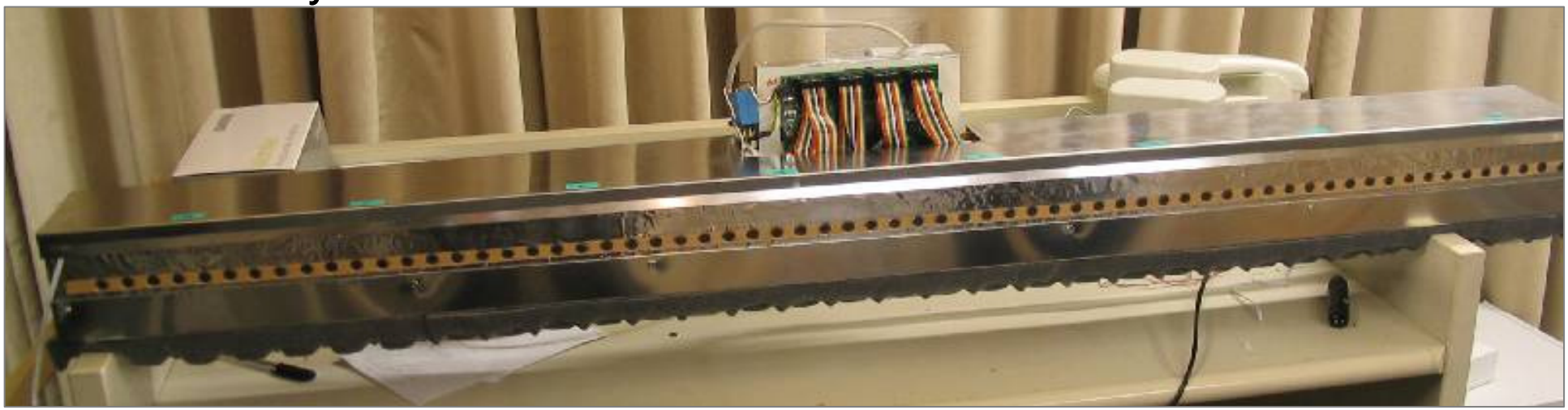
- ~7m x ~6m x 3m size
- Reverberation time ~ 450 ms
- 4 reverse T-shaped arrays, close-talking and 8 table microphones

Linear microphone array installed in the CHIL rooms

Some of the next experiments were conducted by using this device, that is a modified version of the NIST MarkIII microphone array:

- 64 electret microphones
- 2 cm between adjacent microphones (i.e. array length = 126 cm)
- Sampling frequency = 44.1 kHz
- 24 bit precision/sample

→ 8.07 Mbytes/s bandwidth



Details on the modifications designed and implemented at ITC-irst can be found in [Brayda et al.,AES 2005].➡Very good quality of first 16 bits

Outline

- ***Part I:Introduction***
  - The CHIL project: general objectives
  - Foreseen applicative contexts
  - Acoustic scene analysis: distributed microphone networks
- **Part II: Speaker Location**
  - Common approaches and basic techniques
  - Global Coherence Field (GCF) and Oriented GCF (OGCF)
  - Experimental results and NIST benchmarking

Acoustic source location: common approaches and techniques

1. Indirect Methods ⇨ dual-step procedures
2. Direct Methods ⇨ single-step procedures

a. TDOA estimation
b. Apply geometry
(Triangulation, MLE,closed-form, CLS, Sph.Interp., Sph.Inters., etc.)

i. Memoryless solutions ⇨ frame-by-frame independent estimation of the source position

ii. Memory-based solutions ⇨ regularize a sequence of positions on a temporal interval longer than one frame

Ext.Kalman, Particle filtering, Rec. Gauss, etc

Maximization in space of a 3-D (or 2-D) Power (or Coherence) Field function

Eigenanalysis, AEDA, MUSIC, EB-ESPRIT, BSS, etc.

More recent works on high-resolution spectral estimation based on the signal correlation matrix and other techniques

Location of Near-field vs Far-field sources

- **Far-field** source ⟺ Propagation of sound as a plane wave
- **Near-field** source ⟺ Propagation of sound as a spherical wave

A source can be considered to be in the far-field if ⇒ $r > \frac{2L^2}{\lambda}$

where: r is the distance to the array,

L is the length of the array, and

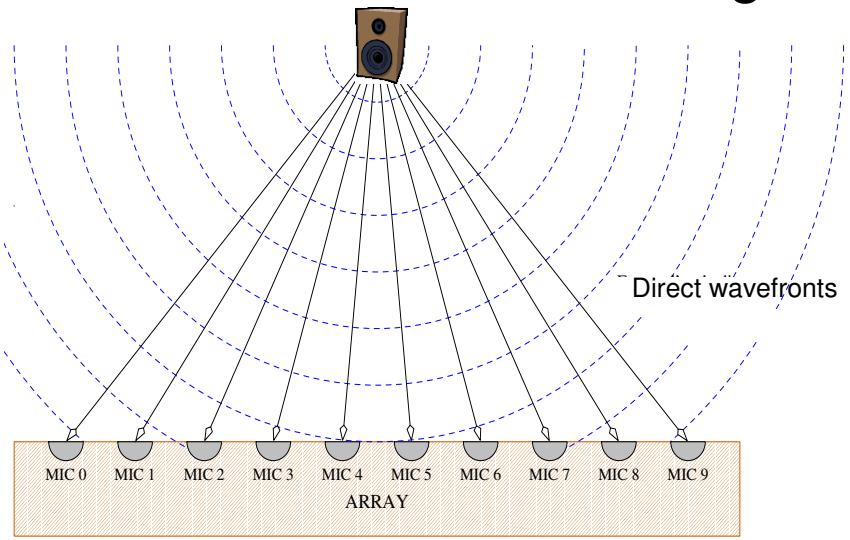
λ is the wavelength of the arriving wave. f =

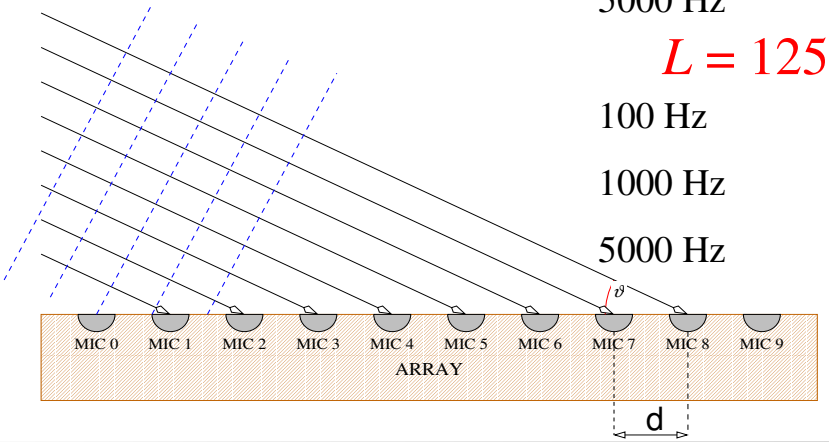
$L = 25\ cm$

100 Hz	$r >\sim$ 3.7 cm
1000 Hz	37 cm
5000 Hz	1.84 m

$L = 125\ cm$

100 Hz	92 cm
1000 Hz	9.2 m
5000 Hz	46 m

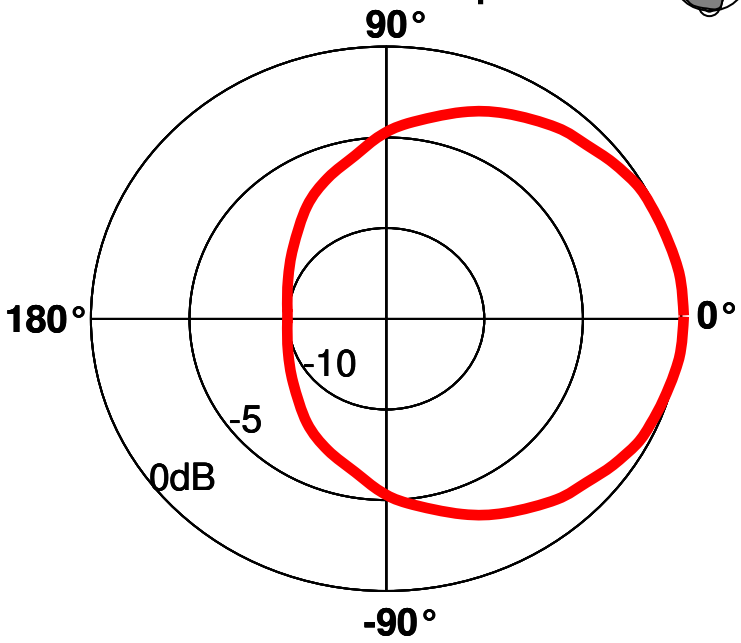




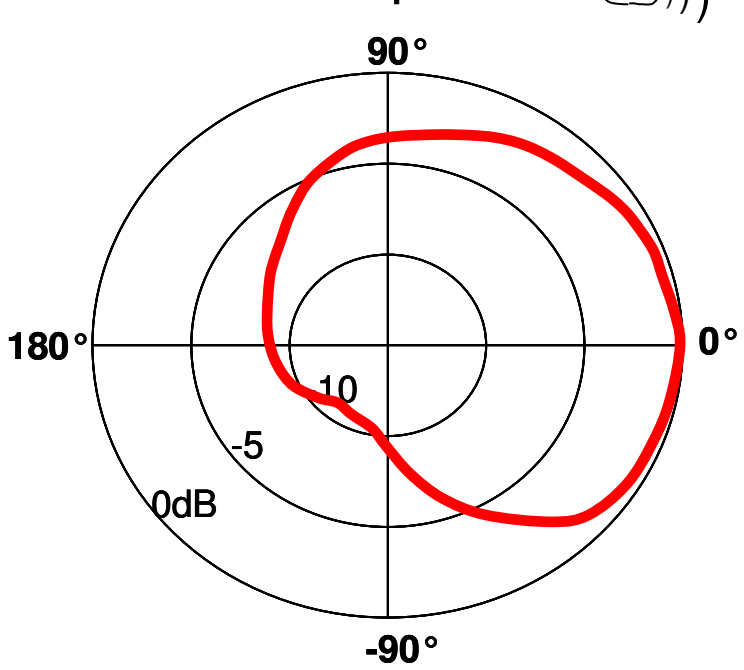
Directivity of speech emission by a talker

- We can not assume the speaker can be modeled as an omnidirectional source.
- If we were able to measure the radiation effects of a talker we would expect to obtain patterns as the following ones:

Horizontal plane



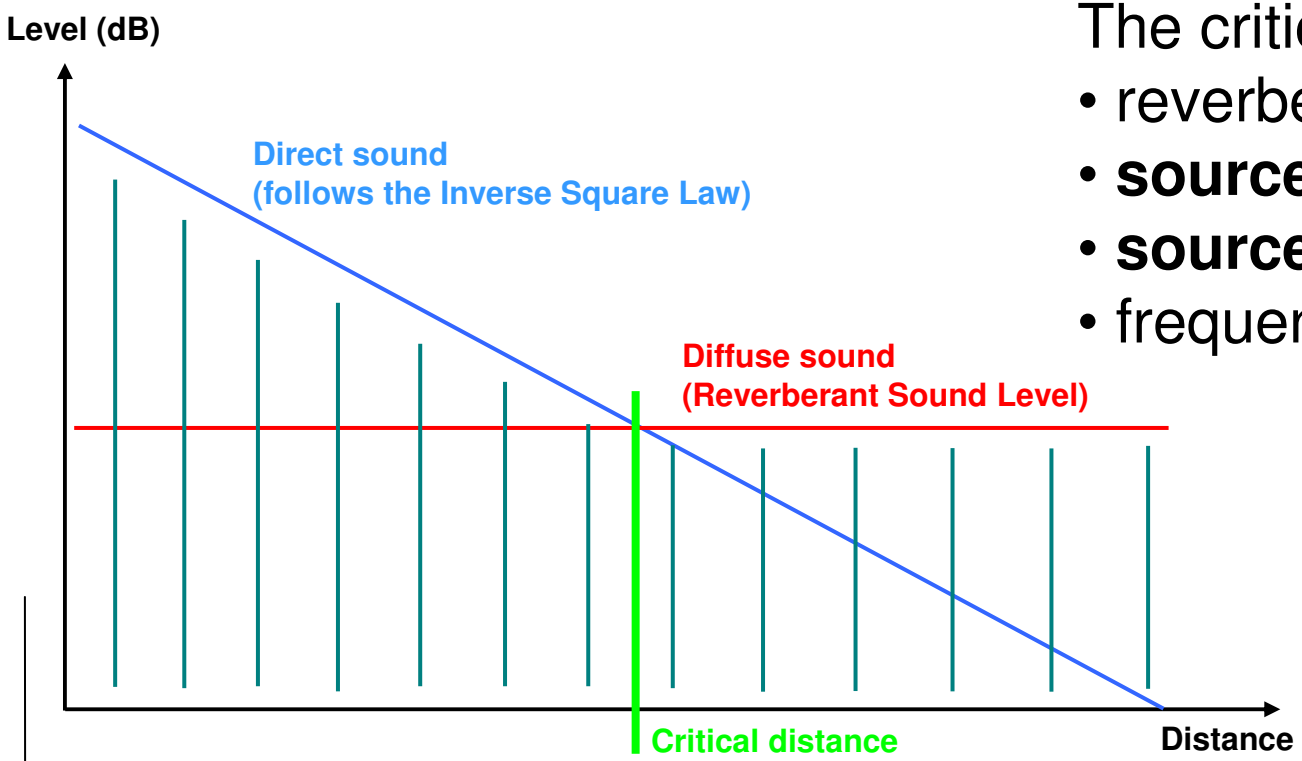
Vertical plane



Critical distance

The ***critical distance***: distance from the sound source at which the direct and reflected sound intensities are equal.

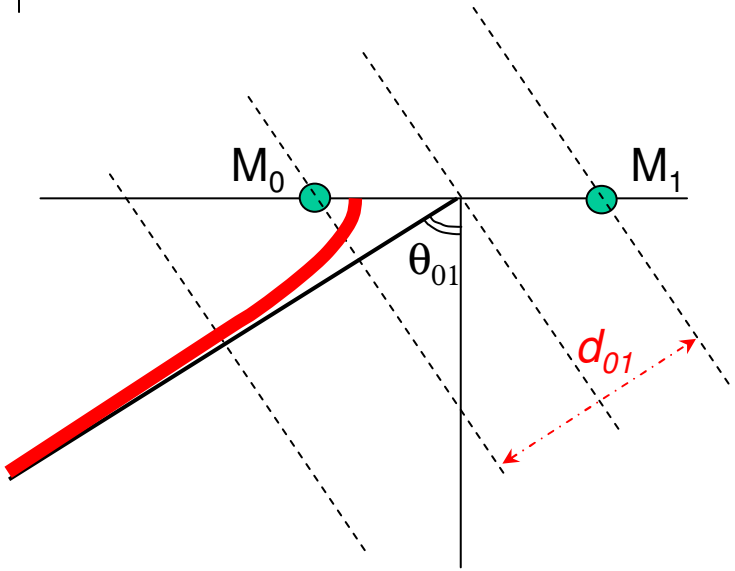
Beyond the radius of critical distance, Signal to Reverberation Ratio (SRR) is negative (except for sound onset)



The critical distance depends on:

- reverberation time
- **source radiation pattern**
- **source orientation**
- frequency

Trivial two-step 2D solution based on two microphone pairs



$$\theta_{01} = arcsin\left(\frac{c\,\tau_{01}}{|m_1 - m_0|}\right)$$

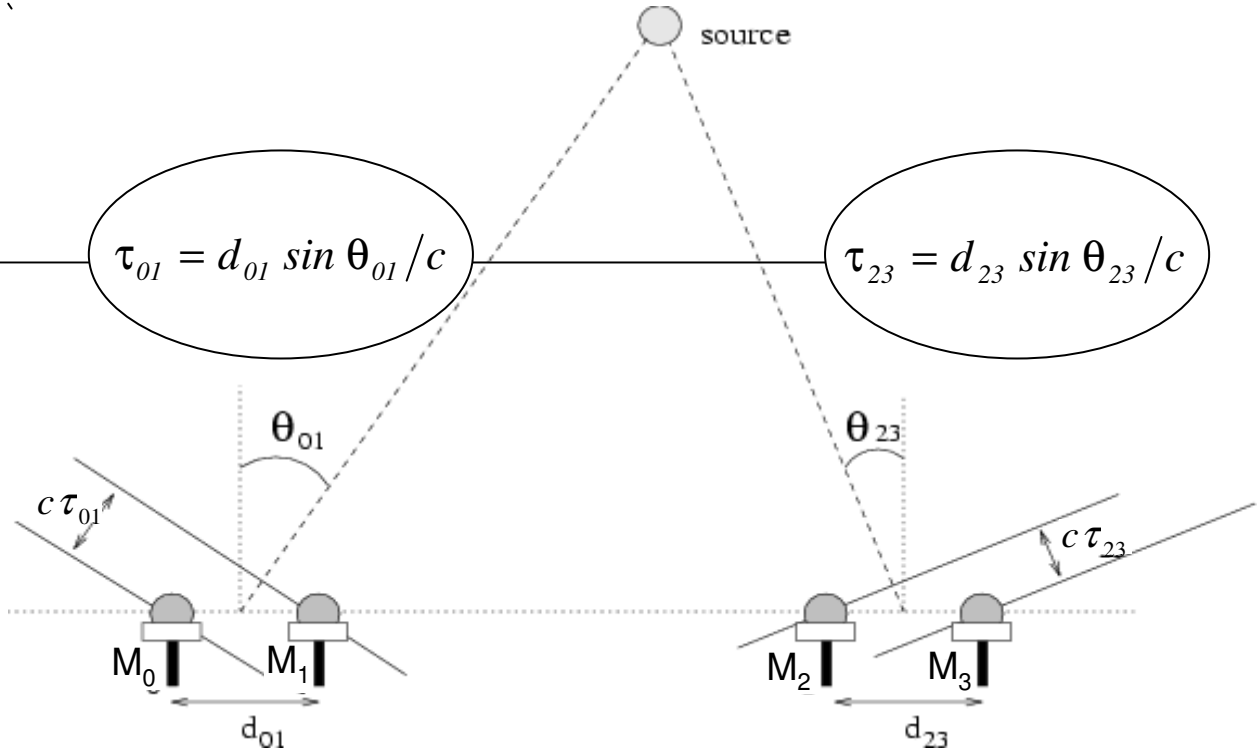
$$d_{01} = c\tau_{01}$$

a. Estimate the two relative delays τ_{01}, τ_{23}

b. Cross the resulting directions

$\tau_{01} = d_{01}\ sin\ \theta_{01}/c$

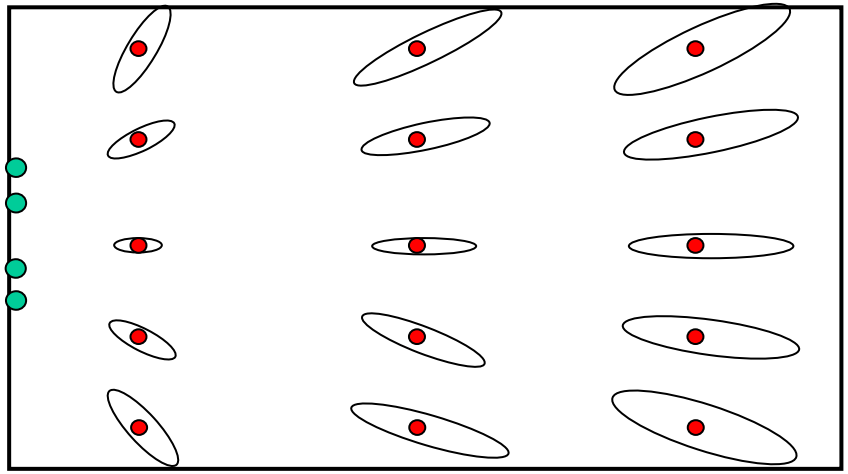
$\tau_{23} = d_{23}\ sin\ \theta_{23}/c$

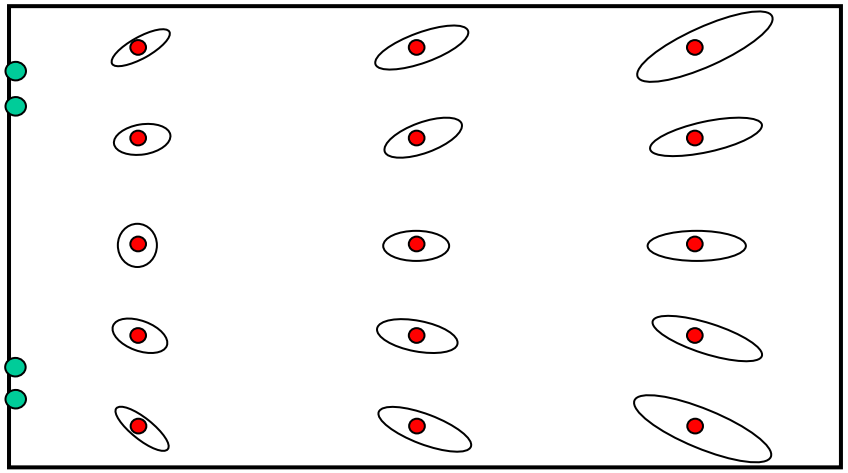


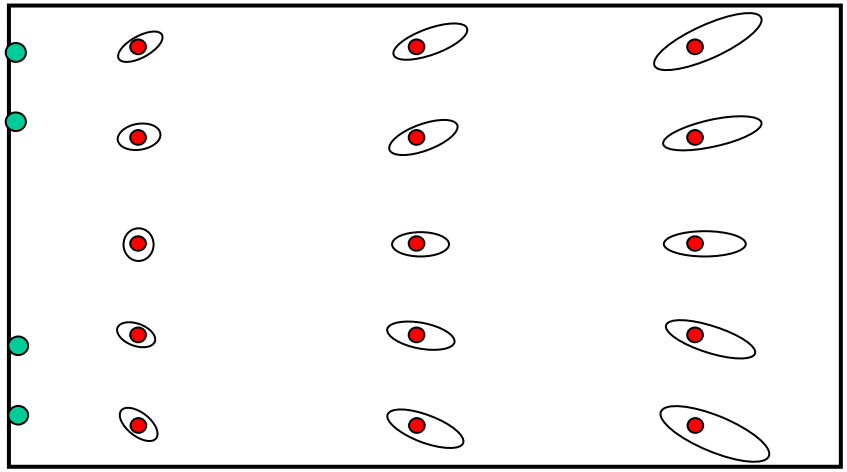
TDOA error statistics vs location accuracy

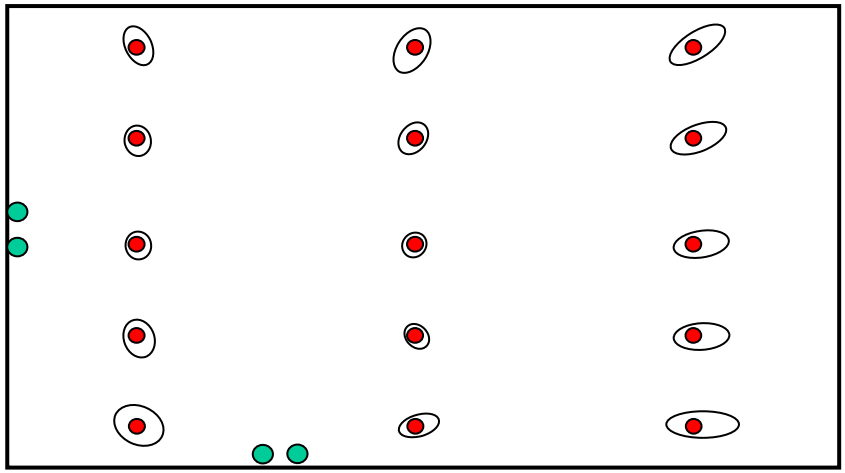
Changing the array geometry vs potential 2D location accuracy

Variance of location estimate changes as a function of the microphone placement









Time Delay Estimation: Cross-Correlation

Estimation of mutual time delay between two signals $y_0(t)$ and $y_1(t)$:

Cross-correlation: $$cc_{01}(\tau) = \int_{-\infty}^{+\infty} y_0(t) y_1(t+\tau) dt$$

in frequency domain: $$cc_{01}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} Y_0(\omega) Y_1^*(\omega) e^{j\omega\tau} d\omega$$

It is estimated for a temporal window centered in t and length T_w

$$cc_{01}(\tau) = \int_{t-T_w/2}^{t+T_w/2} y_0(t) y_1(t+\tau) dt$$

delay estimate as peak of cross-correlation: $$\hat{\tau}_{01} = \arg\max_{|\tau|<d/c} [cc_{01}(\tau)]$$

The peak of cross-correlation is influenced by the signals' autocorrelation.It may be quite broad and sensitive to noise and reverberation.

Generalized Cross Correlation (GCC)

A sharper peak can be obtained by prefiltering of the signals.

Generalized Cross-Correlation (Knapp-Carter'76):

$$gcc_{01}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}[G_0(\omega)Y_0(\omega)][G_1(\omega)Y_1(\omega)]^* e^{j\omega\tau}d\omega$$

$$gcc_{01}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\Psi_{01}(\omega)\cdot[Y_0(\omega)Y_1^*(\omega)]\, e^{j\omega\tau}d\omega$$

The **GCC- PHAT** (Phase Transform), corresponding to the **CSP** (Cross-power Spectrum Phase) analysis, is obtained with:

$$\Psi_{01}(\omega) = \frac{1}{|Y_0(\omega)Y_1^*(\omega)|} \Rightarrow$$

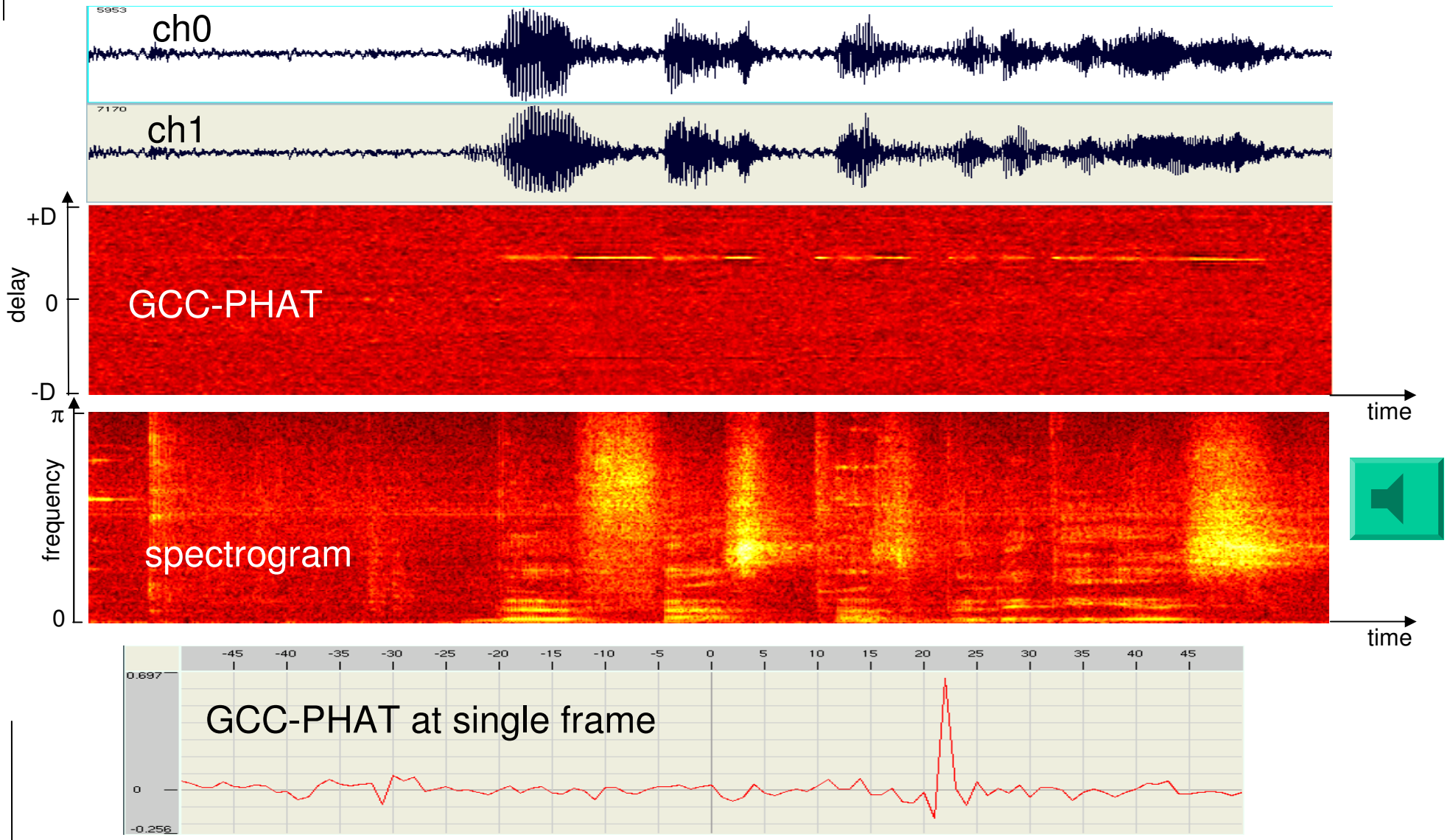
$$\Rightarrow gcc_{01}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{Y_0(\omega)Y_1^*(\omega)}{|Y_0(\omega)Y_1^*(\omega)|}e^{j\omega\tau}d\omega$$

Amplitude is normalized to 1. Only phase information is preserved!

$$\Rightarrow \hat{\tau}_{01} = \arg\max_{|\tau|<d/c}[gcc_{01}(\tau)]$$

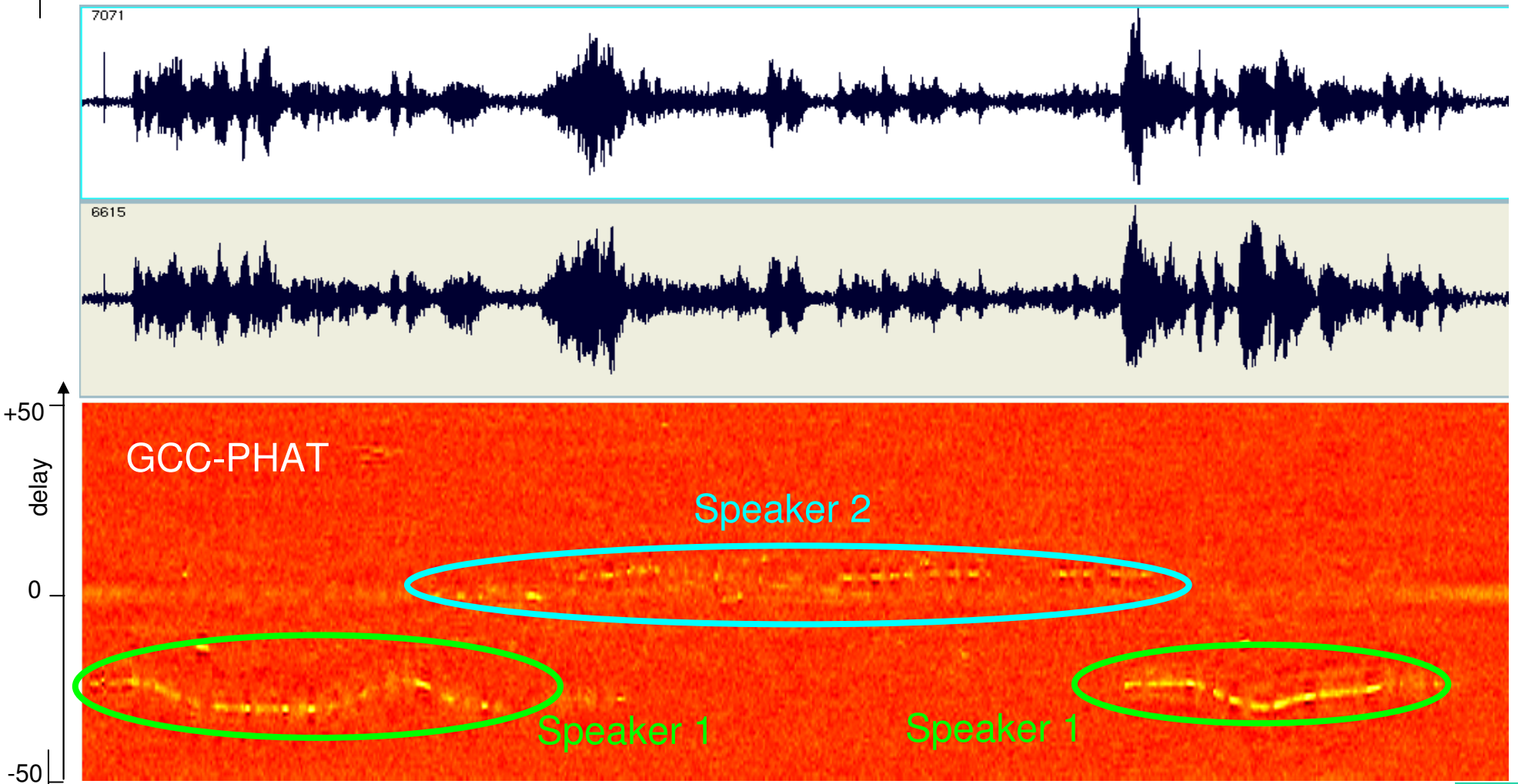
An interpolation refinement is needed to get accurate fractional delay estimates.

Application of GCC-PHAT analysis: single speaker

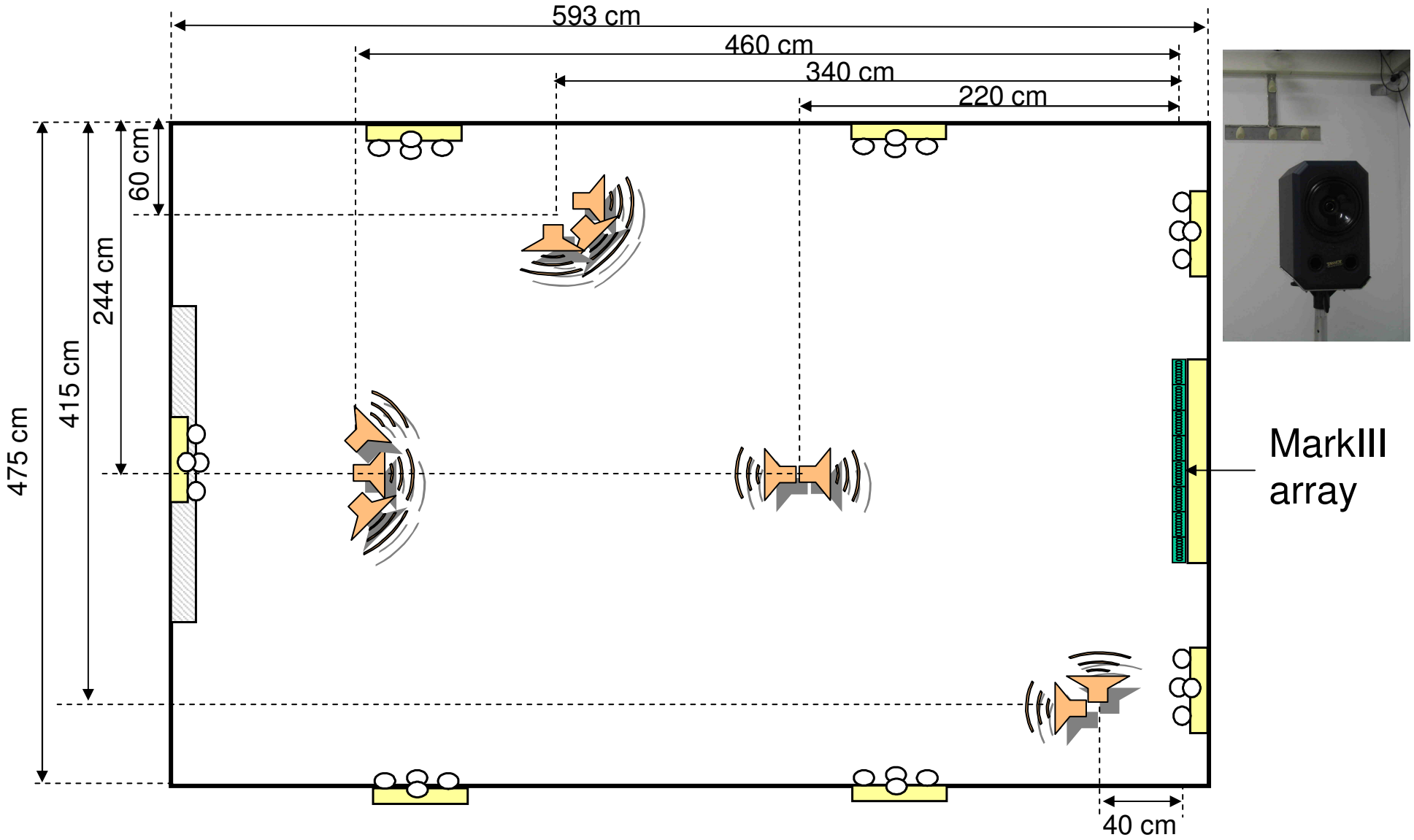


Maurizio Omologo – omologo@itc.it

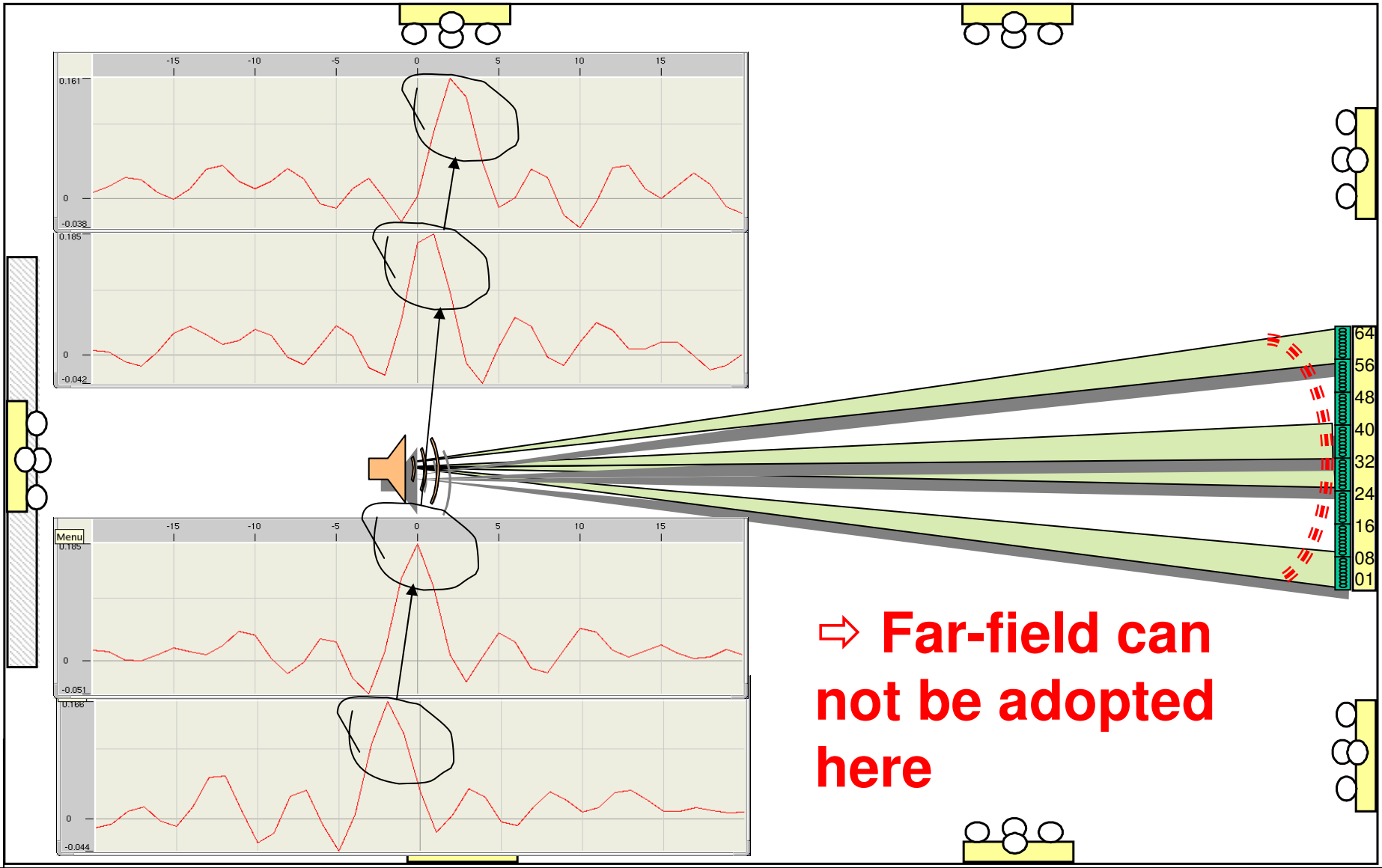
Application of GCC-PHAT analysis: two competitive speakers



Real experiments in the CHIL room at ITC-irst



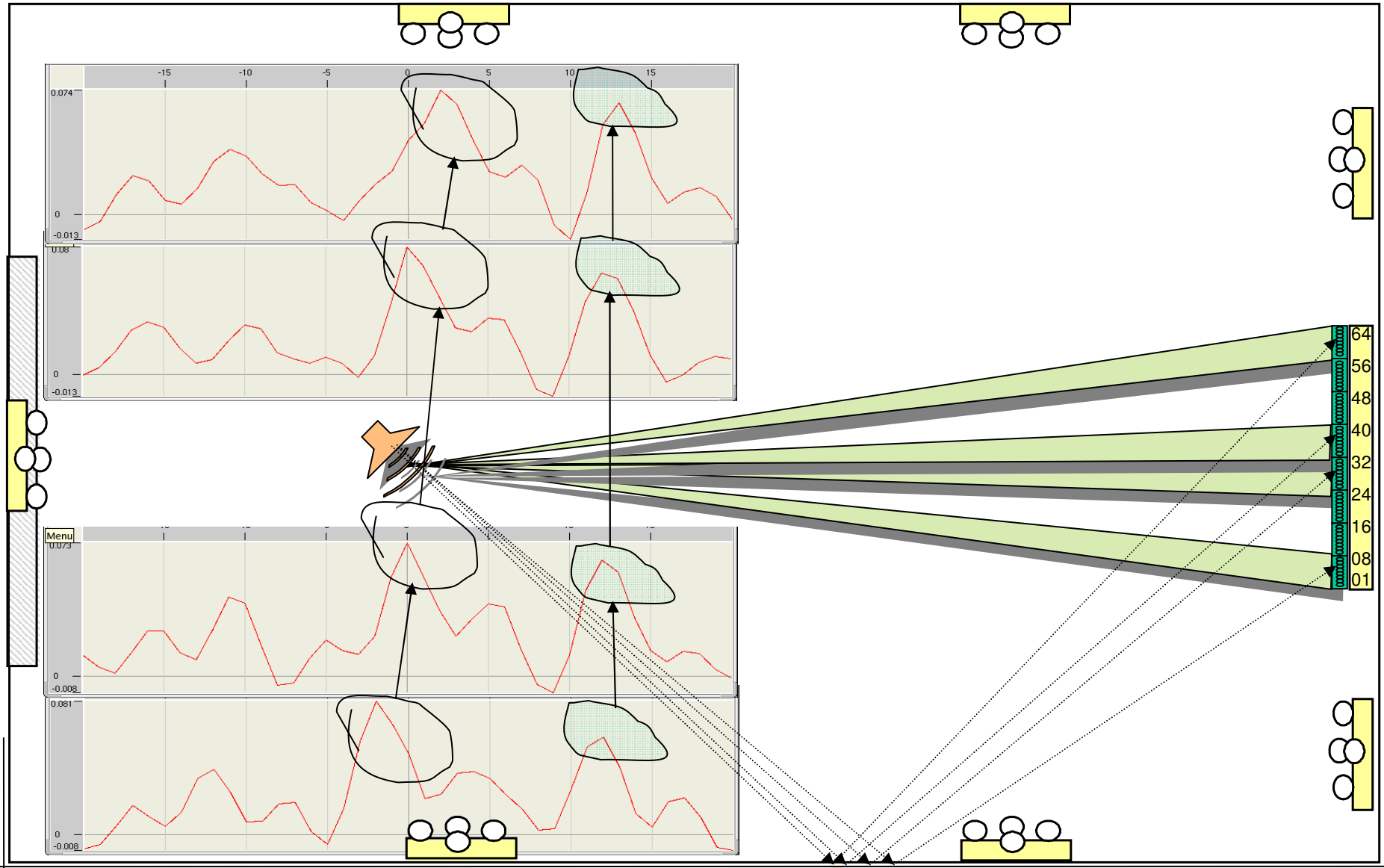
Loudspeaker oriented towards the array



⇨ Far-field can not be adopted here

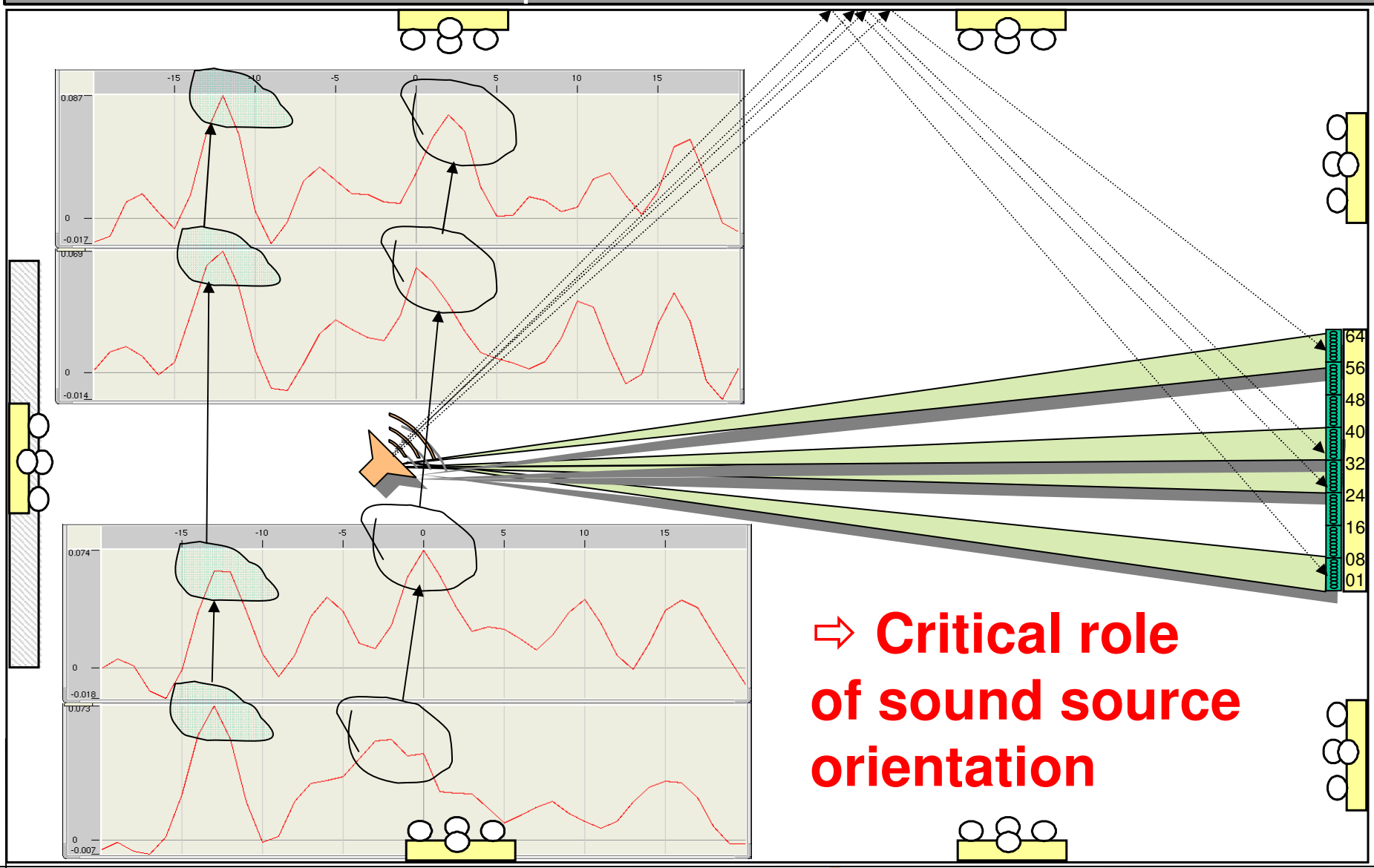
Maurizio Omologo – omologo@itc.it

Loudspeaker oriented to the right



Maurizio Omologo – omologo@itc.it

Loudspeaker oriented to the left



⇨ **Critical role of sound source orientation**

Maurizio Omologo – omologo@itc.it

Coherence with diffuse noise

Complex Coherence: $$\gamma_{xy}(f) = \frac{P_{xy}(f)}{\sqrt{P_{xx}(f)P_{yy}(f)}}$$

Magnitude Squared Coherence (MSC): $$C_{xy}(f) = \frac{\left|P_{xy}(f)\right|^2}{P_{xx}(f)P_{yy}(f)}$$

P_{xx}, P_{yy} Power spectra

P_{xy} Cross spectrum

can be obtained by Welch's method of power spectrum estimation

$0 \le C_{xy}(f) \le 1$ provides a measure of the correlation of the signals acquired by two different microphones

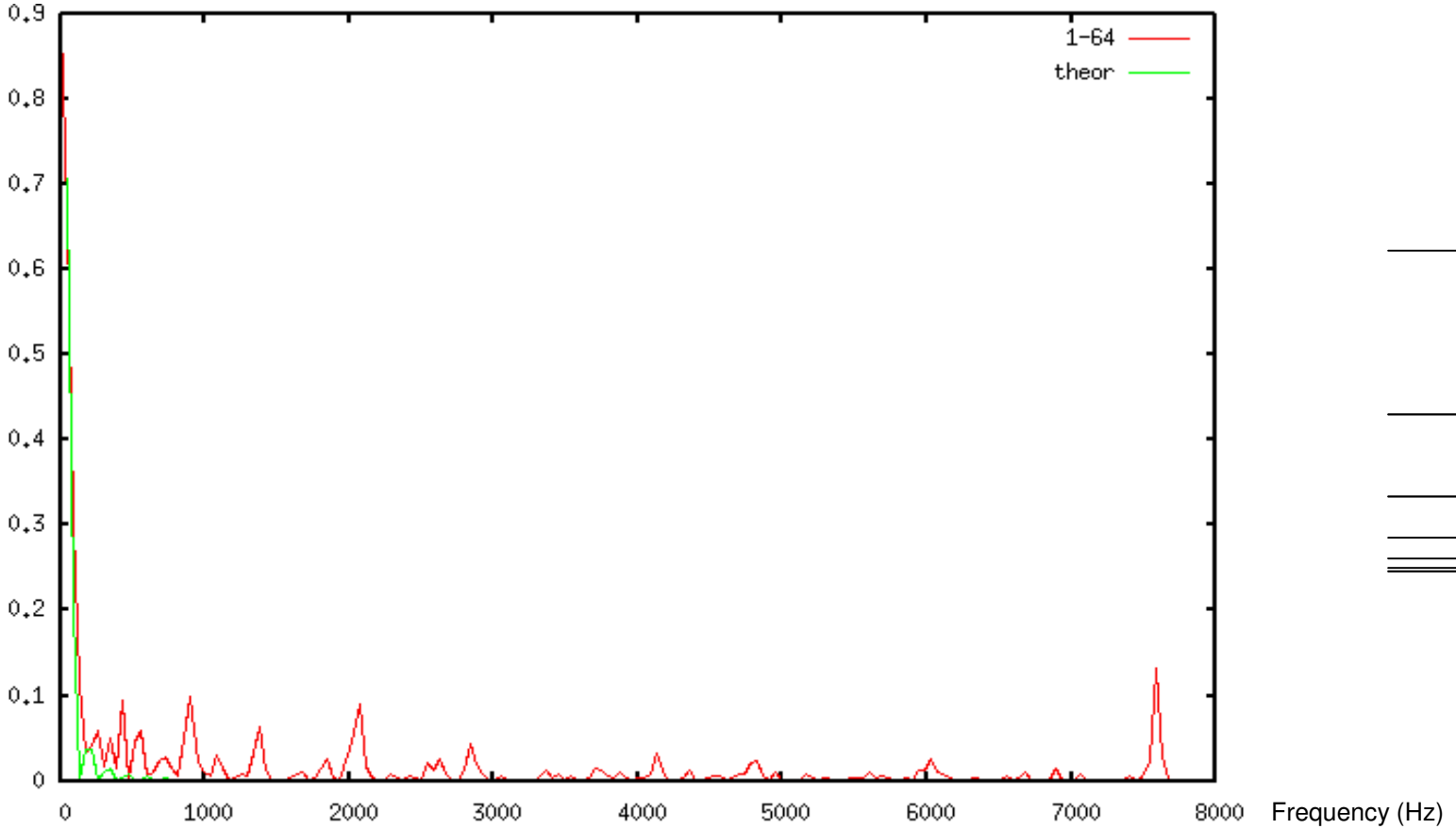
For perfectly diffuse noise (spherically isotropic) and omnidirectional microphones, the theoretical MSC function is:

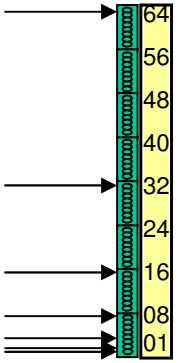
[see chapter 4, G. Elko, in Brandstein-Ward]

$$C_{xy}(f) = \left(\frac{sin(2\pi f \cdot d/c)}{2\pi f \cdot d/c} \right)^2$$

Experiments in the CHIL room at ITC-irst

$C_{xy}(f)$



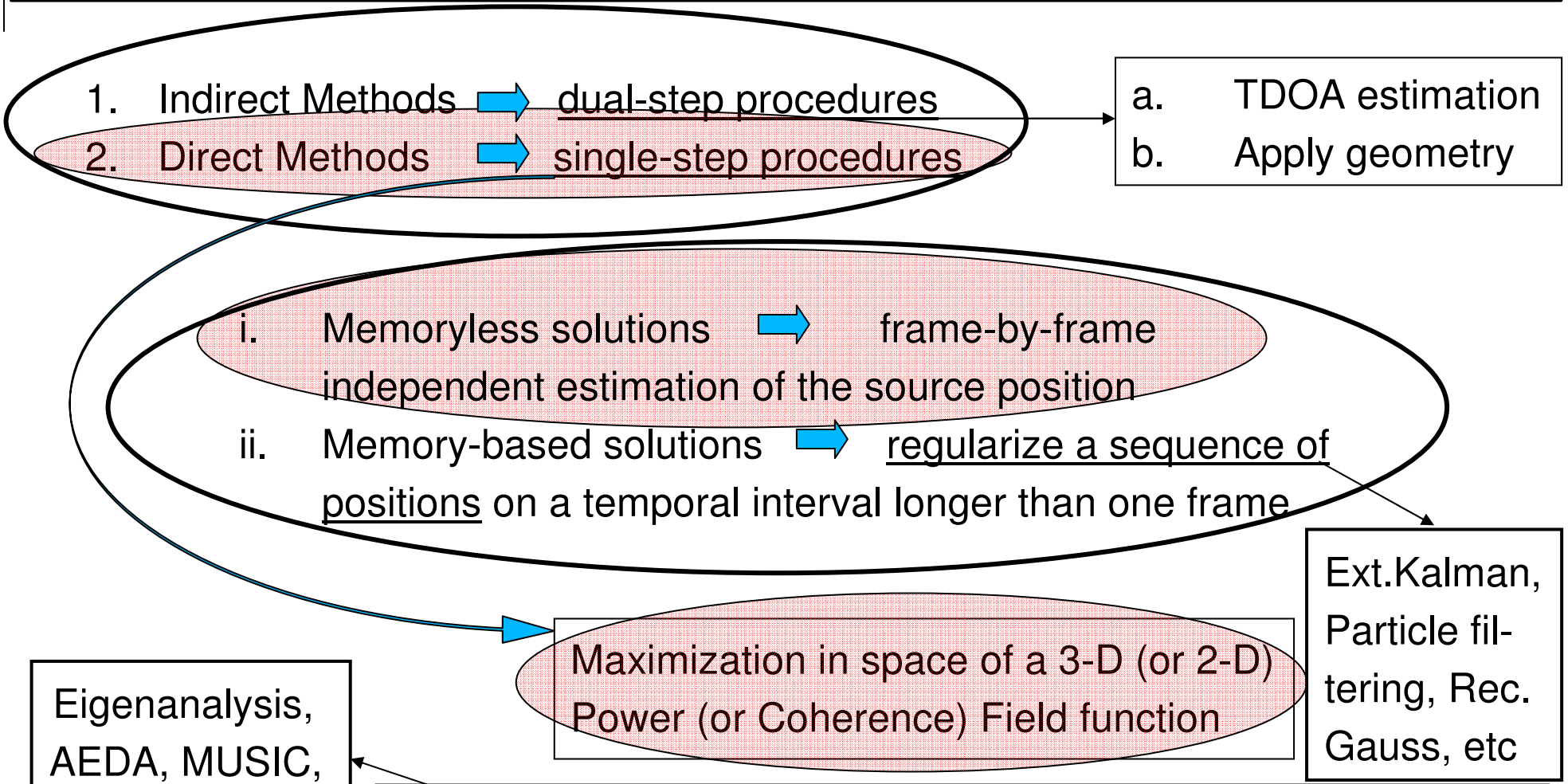


Analysis of the spatial coherence of background noise by means of different microphone pairs of the Mark III array, in a reverberant environment.

Comparison with the theoretical behavior for perfectly diffuse noise.

Good clues to classify the noise diffuseness in the environment

Acoustic source location: common approaches and techniques



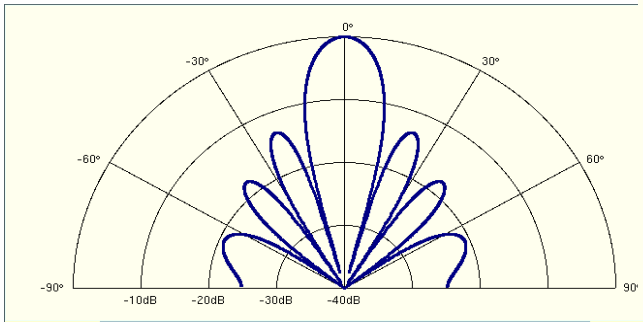
Power Field/Steered Response Power

Given the ***Delay-and-Sum beamforming***, i.e.

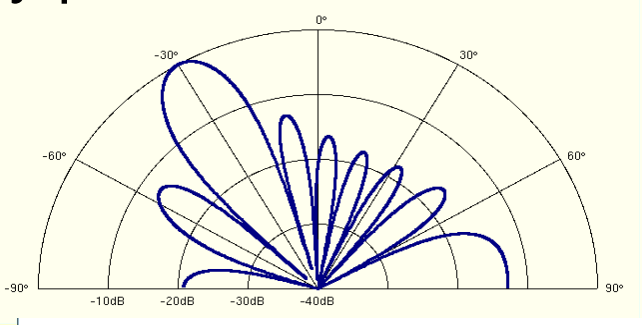
$$z(t,\mathrm{s}) = \sum_{n=0}^{M-1} y_n\left(t + T_n(\mathrm{s})\right)$$

T_n=steering delays

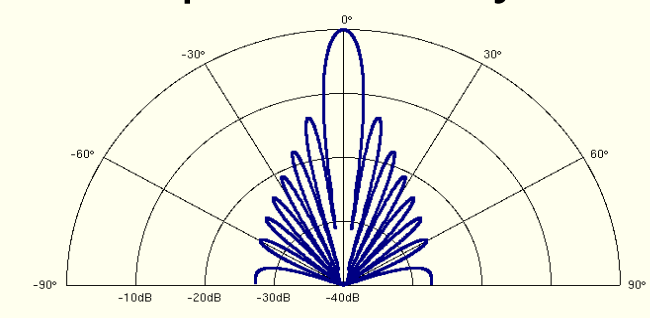
Possible directivity patterns with a linear microphone array:



Array steered at 0° (1kHz)



Array steered at -30° (1kHz)



Array steered at 0° (2kHz)

- D&S Beamforming can be used to "scan" the space and look for a maximum of received acoustic power. This is the ***PF*** (***Power Field***) approach [Alvarado 1990], recently renamed as ***SRP*** (***Steered Response Power***).
- Drawbacks:
  - computationally expensive
  - no strong global peak
  - highly dependent on the spectral content of the signals

Global Coherence Field

- A hybrid approach that conjugates the advantages of TDOA method and of Power Field.

Given a set M_P of microphone pairs the Global Coherence Field* (GCF) [Omologo-Svaizer 1993, 1997] is computed at time instant t as:

$$GCF(t,\mathbf{s}) = \frac{1}{M_P} \sum_{(i,k)\in\{M_P\}} gcc_{ik}(t, \delta_{ik}(\mathbf{s}))$$

where $\delta_{ik}(\mathbf{s})$ denotes the theoretical delay for the (i,k) microphone pair having assumed that the source is in position $\mathbf{s}$.

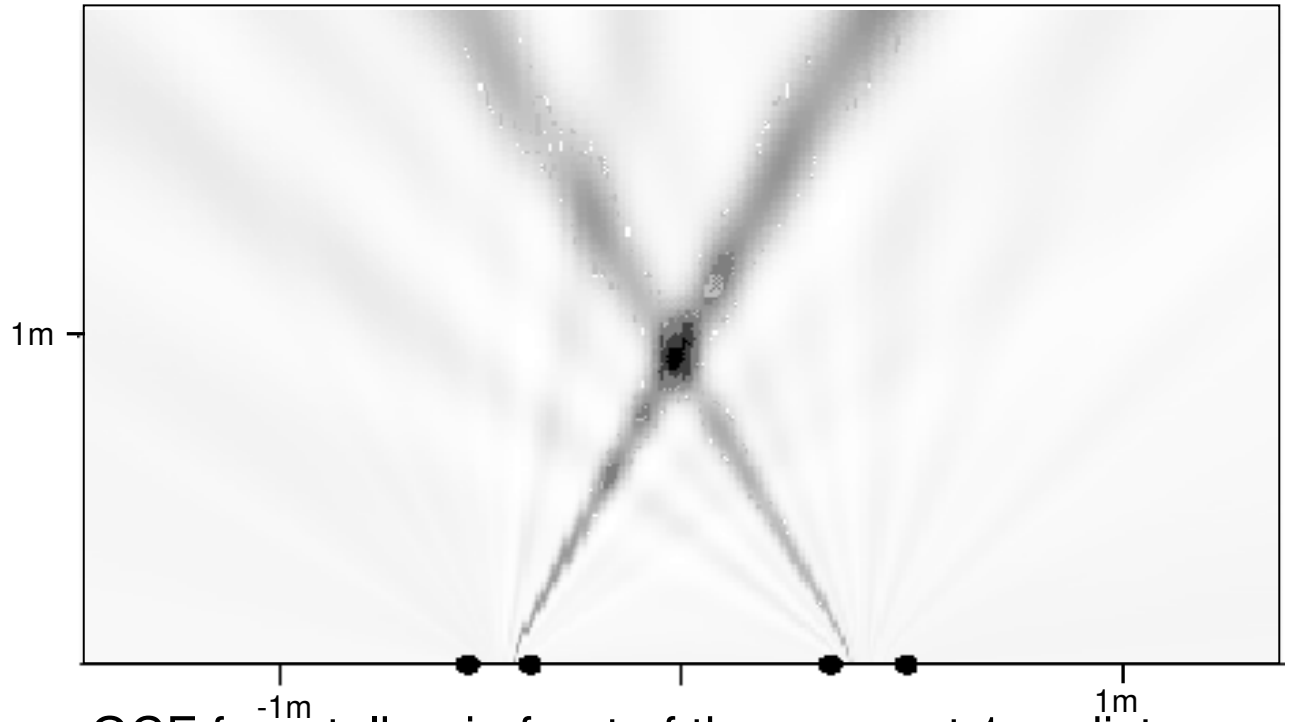
$$\Rightarrow \quad \hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s}} [GCF(t,\mathbf{s})]$$

- Advantages: In GCF acoustic maps, peaks are sharper than in Power Field maps, with a consequent decreased sensitivity to noise and reverberation. Moreover, it is a direct single-step method.

*** Other recent literature [see Brandstein-Ward 2001] uses more commonly the term SRP-PHAT to indicate the above described GCF technique.**

GCF-based acoustic maps

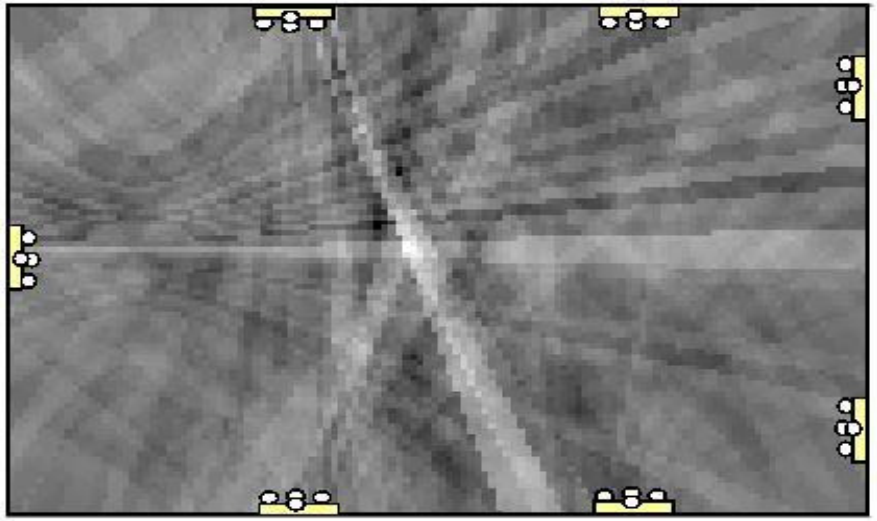
Example of Global Coherence Field accounting for the contributes of the CSP functions related to two microphone pairs:

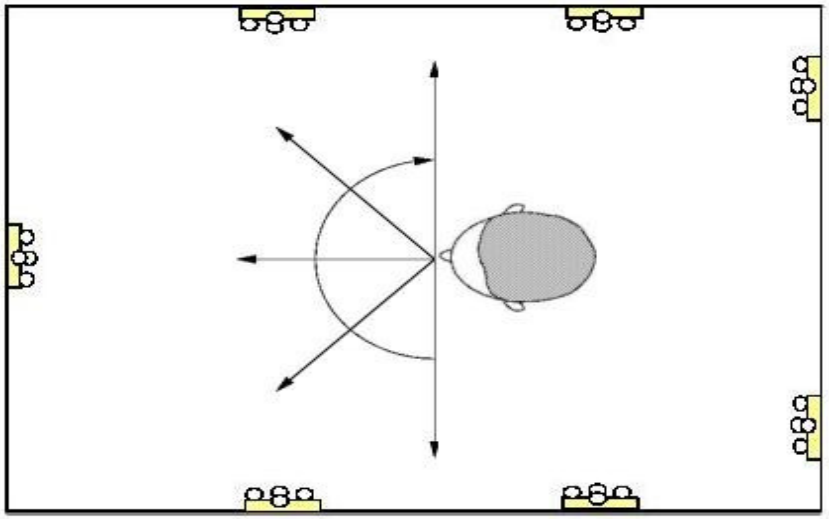


GCF for a talker in front of the array, at 1 m distance

Note that GCF preserves all the information expressed by the CSP functions, not only the bearing directions associated to main peaks.

Use of GCF to estimate Head Orientation





Example of 2D GCF in the CHIL room at ITC-irst

(Note: darker points here denote low values)

According to head orientation the contribution of the various micro-phone pairs have different strength

➡ The relative variations of GCF around the source position are clues to deduce source orientation

➡ The audio map of GCF can be exploited to derive information about talker orientation

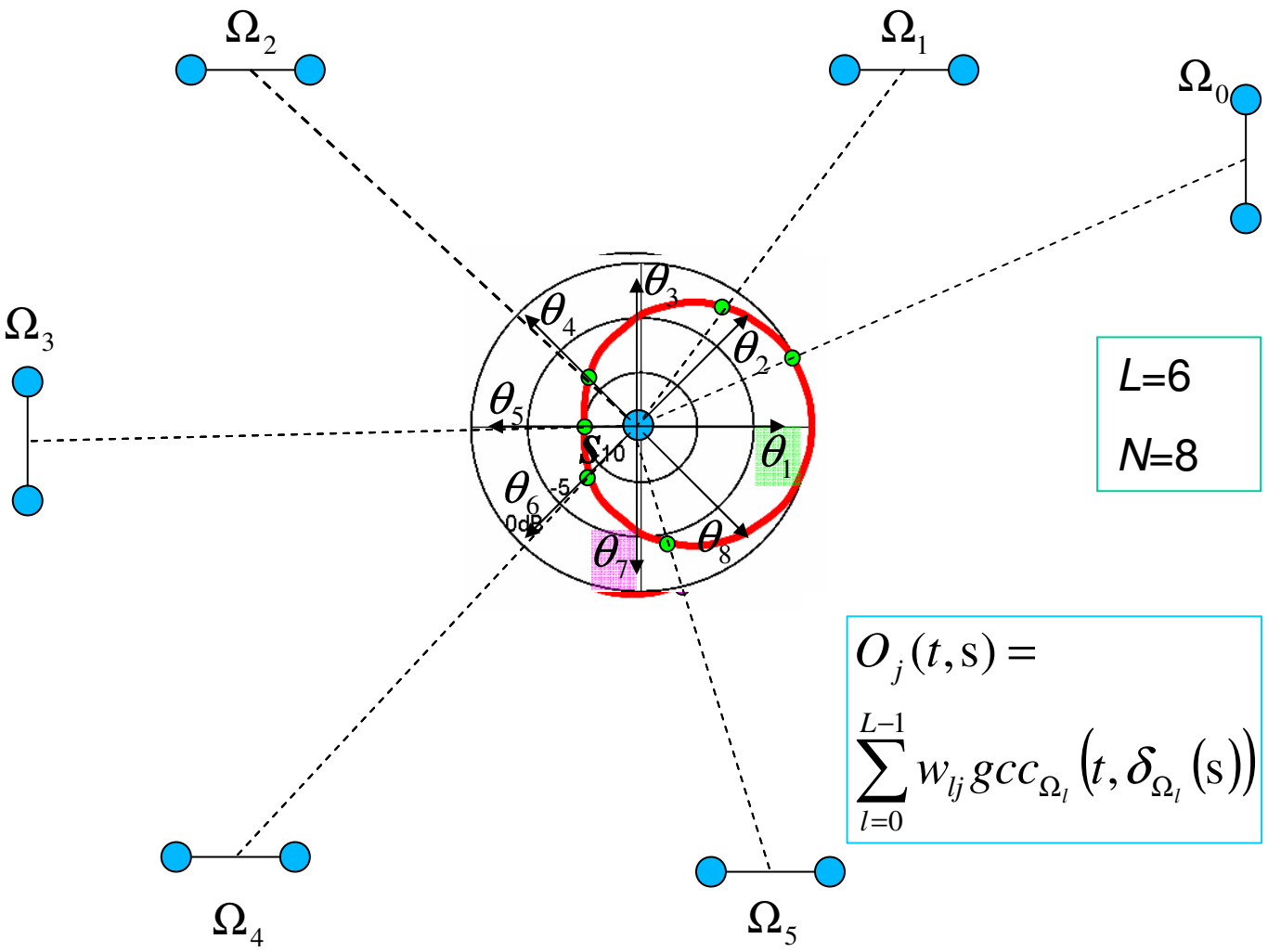
➡ ***Oriented Global Coherence Field*** (one GCF for each direction)

Oriented Global Coherence Field (OCGF): basic concept

Given:

1) A set of *L* microphone pairs Ω_l
2) A generic point ***s***
3) A direction *j* of a given set of N possible directions

- Compute a gcc_{Ω_l} for each pair at the angle steering to the point ***s***
- Weight each gcc_{Ω_l} according to the value w_{lj} assumed by a polar function oriented at the selected direction θ_j



L=6

N=8

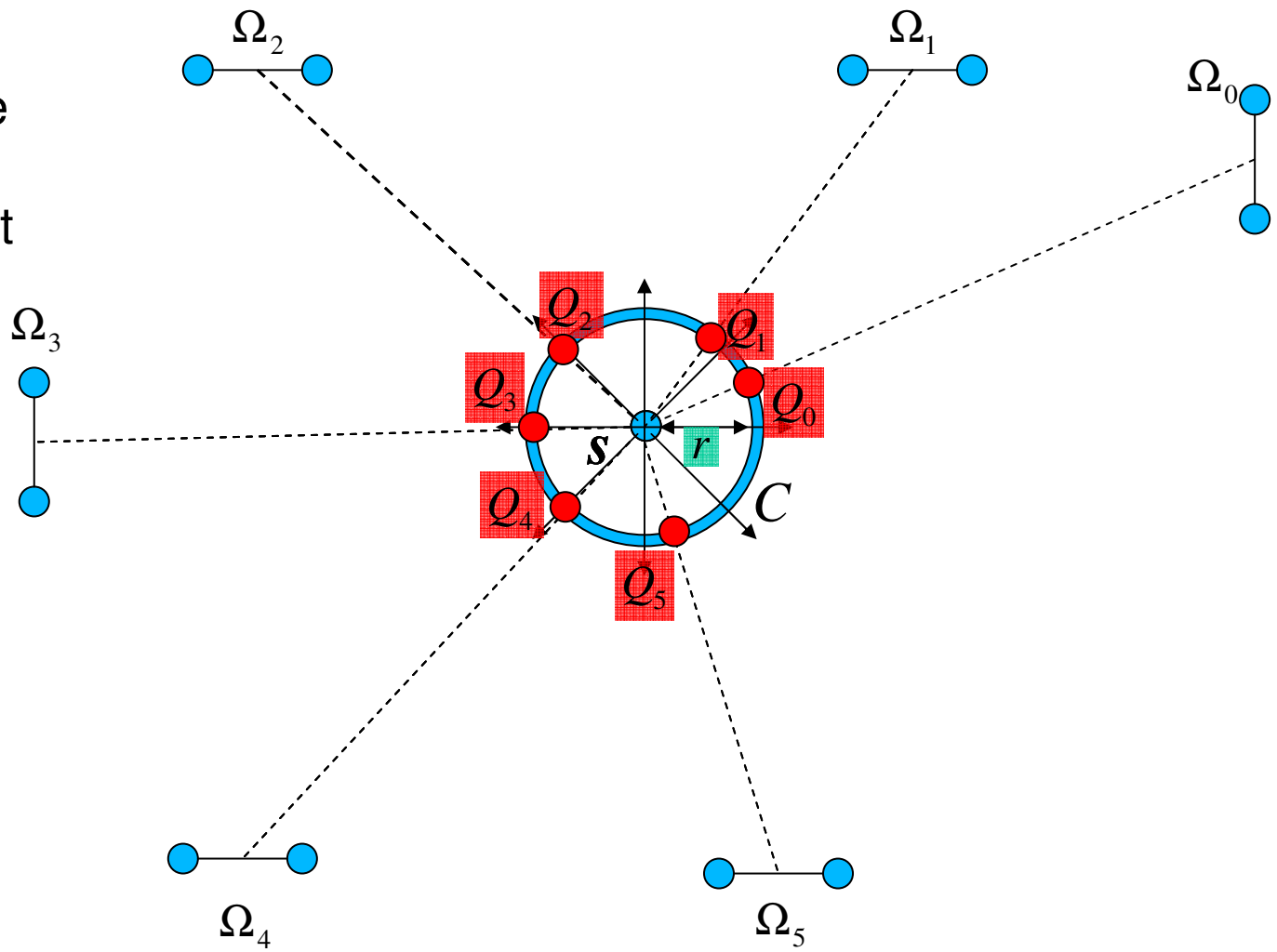
$$O_j(t,s) = \sum_{l=0}^{L-1} w_{lj}\, gcc_{\Omega_l}\left(t, \delta_{\Omega_l}(s)\right)$$

Oriented Global Coherence Field (OGCF): procedure

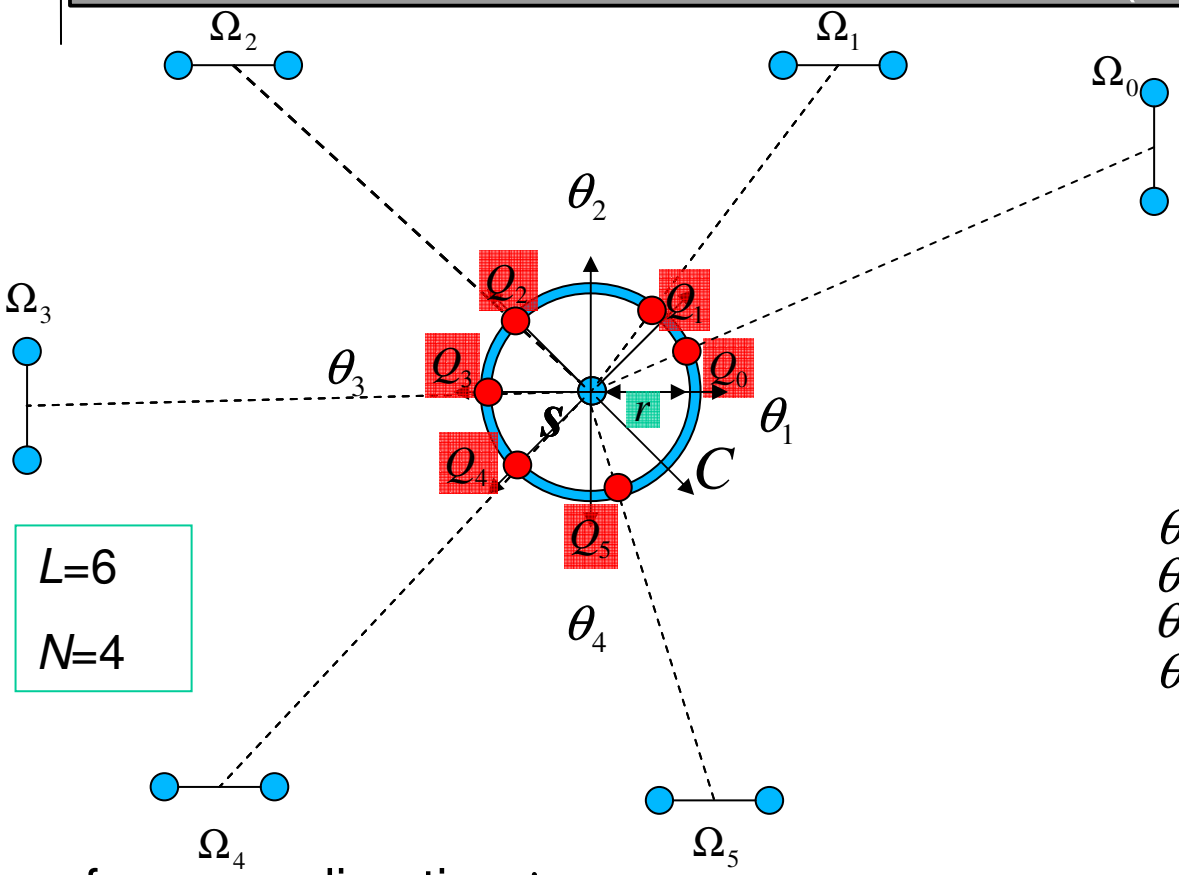
Given:

1) a set of *L* microphone pairs Ω_l
2) a **circle** *C* centered at the given point ***s*** and having a ***radius*** *r*
3) N possible directions

Consider the intersection points Q_l between the lines from **s** to Ω_l and the selected direction *j*

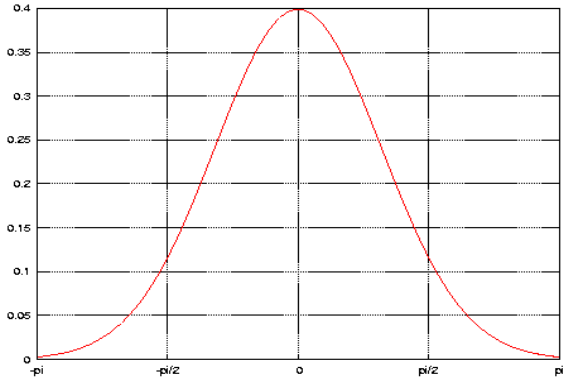


Oriented Global Coherence Field (OGCF): weighting function



L=6

N=4



θ_1	Q_3	Q_4	Q_5	Q_0	Q_1	Q_2
θ_2	Q_5	Q_0	Q_1	Q_2	Q_3	Q_4
θ_3	Q_0	Q_1	Q_2	Q_3	Q_4	Q_5
θ_4	Q_2	Q_3	Q_4	Q_5	Q_0	Q_1

Reasonable solution: $w(\Delta\theta)$ is a weight computed from a gaussian function and defined in $[-\pi,\pi]$.

for every direction *j*:

$$O_j(t,\mathrm{s}) = \sum_{l=0}^{L-1} w_{lj} GCF_{\Omega_l}(t, Q_l)$$

$$\hat{s}(t) = \arg\max_{s,j} O_j(t,s)$$

TDOA-based systems evaluated in the NIST'05 SLOC evaluation

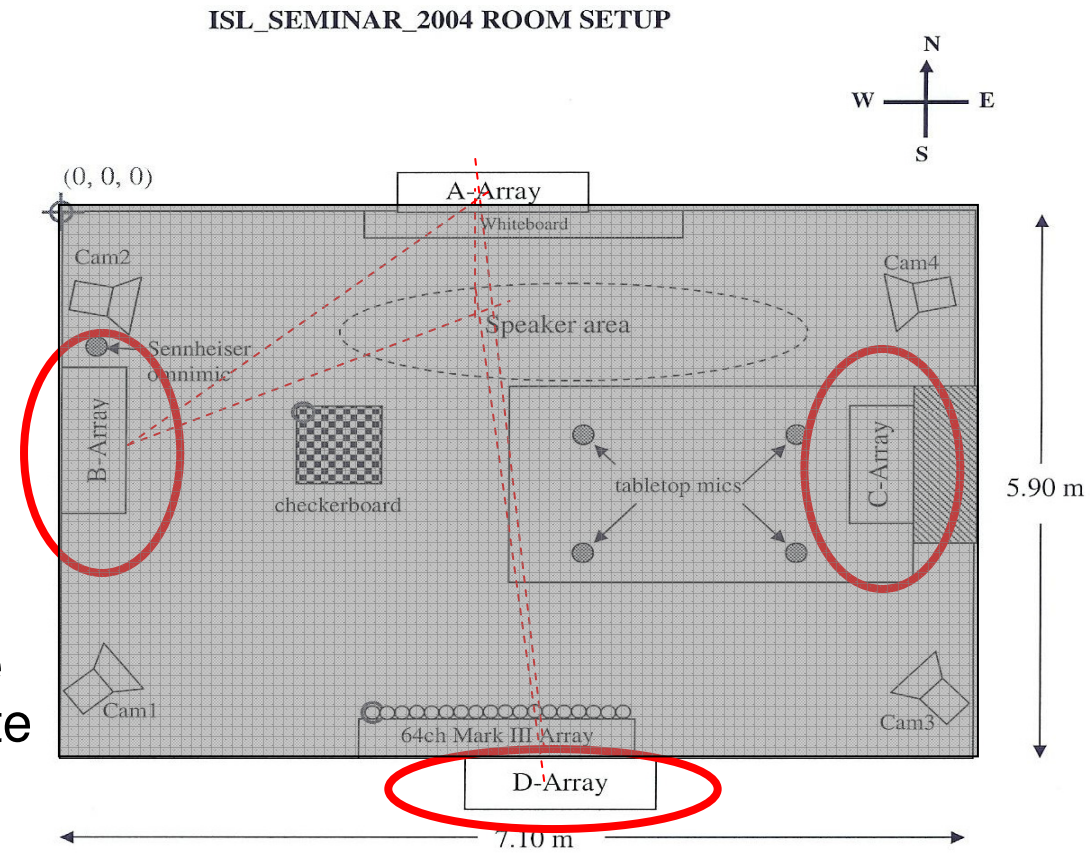
Method 1, based on a **two-step procedure**:

1a) Use of two T-shaped arrays (B and D), two pairs for 2D(x-y) location

1b) Use of two pairs for the z-coordinate: directions derived by CSP (GCC-PHAT) TDOA analysis

Method 2, based on a **single-step 2D location procedure**

2) Use of three T-shaped arrays and of Global Coherence Field (GCF). Same procedure as in 1b) to derive z-coordinate



Evaluation of SLOC systems

o Seminar segments:

- 13 seminars recorded between November 2004 and February 2005 at Karlsruhe University.
- The material was used for NIST SLOC benchmarking 2005 and as development set for CLEAR SLOC benchmarking 2006.

o Evaluation results:

(Evaluation software and criteria are introduced in [Omologo et al. 2006] and in: *http://www.nist.gov/speech/test/rt/rt2005/spring/sloc/CHIL-IRST_SpeakerLocEval-V5.0-2005-01-18.pdf*)

	Technique (SAD threshold)	Output Rate [1/s]	FA Rate [%]	Del. Rate [%]	Loc. Rate [%]	RMSE [mm]	fine RMSE [mm]	Bias [mm]
Method 1	TDE	2.25	42	41	95	309	203	(59,-78,-41)
Method 2	GCF(0)	6.21	81	7	87	479	226	(43,-64,-77)
	GCF(0.38)	1.94	39	48	92	327	198	(40,-47,-51)
	GCF(0.75)	0.07	03	96	91	238	159	(80,-22,-57)
Method 3	OGCF(0.15)	5.09	68	13	95	298	193	(-1,-7,-55)
	OGCF(0.20)	3.91	55	23	95	272	193	(-12,-10,-47)
	OGCF(0.25)	2.84	44	36	95	266	192	(-23,-10,-41)
	OGCF(0.30)	2.01	33	50	95	249	191	(-37,-14,-33)

Evaluation of SLOC systems

Technique (SAD threshold)	Output Rate [1/s]	FA Rate [%]	Del. Rate [%]	Loc. Rate [%]	RMSE [mm]	fine RMSE [mm]	Bias [mm]
TDE	2.25	42	41	95	309	203	(59,-78,-41)
GCF(0)	6.21	81	7	87	479	226	(43,-64,-77)
GCF(0.38)	1.94	39	48	92	327	198	(40,-47,-51)
GCF(0.75)	0.07	03	96	91	238	159	(80,-22,-57)
OGCF(0.15)	5.09	68	13	95	298	193	(-1,-7,-55)
OGCF(0.20)	3.91	55	23	95	272	193	(-12,-10,-47)
OGCF(0.25)	2.84	44	36	95	266	192	(-23,-10,-41)
OGCF(0.30)	2.01	33	50	95	249	191	(-37,-14,-33)

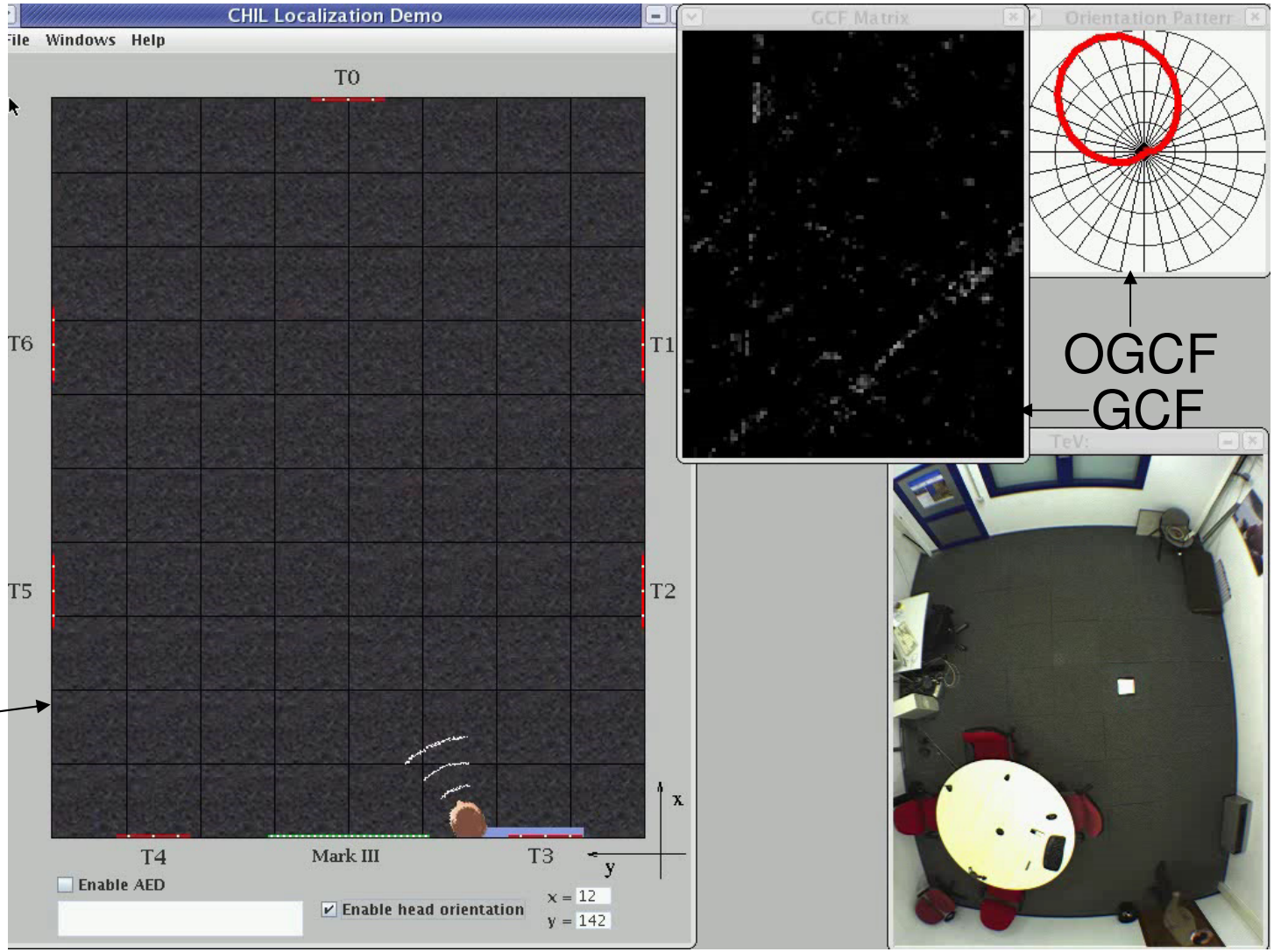
Method 1

Method 2

Method 3

Demo 1: Speaker Tracking and Head Orientation

- 2-D real-time speaker tracking based on 7 microphone pairs
- OGCF location algorithm
- OGCF-threshold based speech activity detection



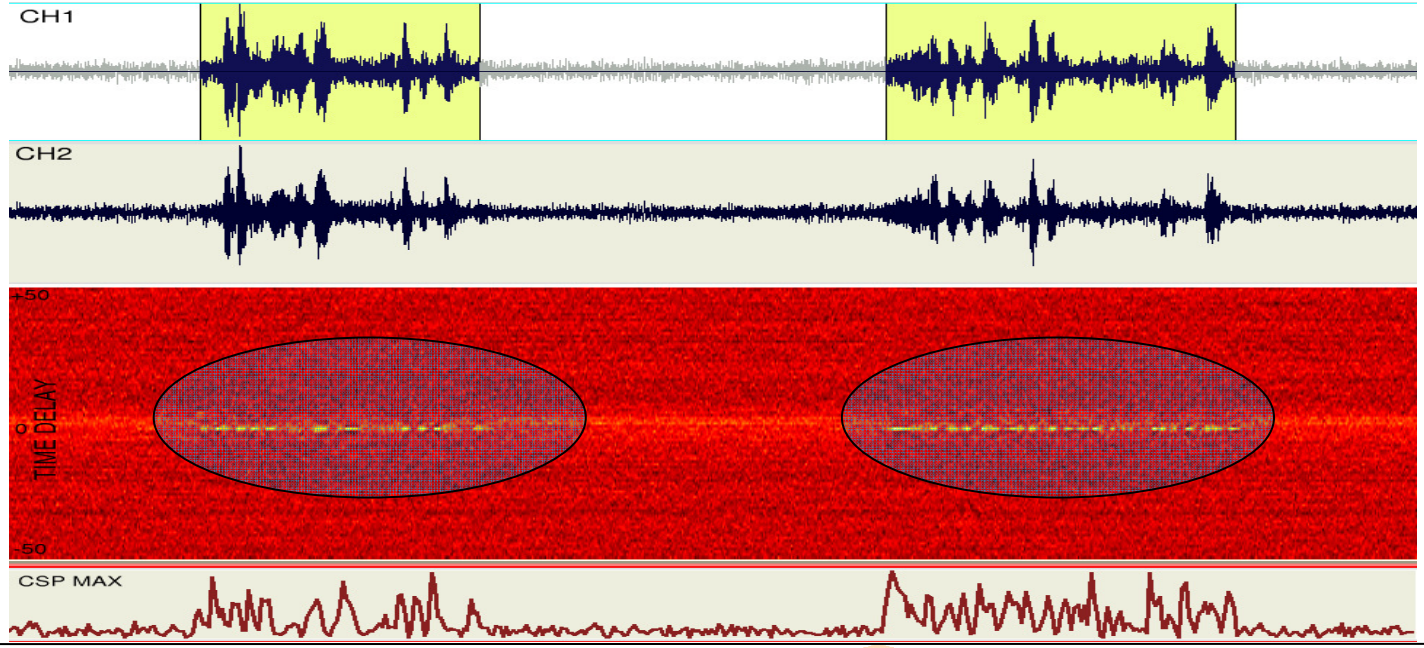
Map of the CHIL room at ITC-irst

Outline

- ***Part I:Introduction***
  - The CHIL project: general objectives
  - Foreseen applicative contexts
  - Acoustic scene analysis: distributed microphone networks
- **Part II: Speaker Location**
  - Common approaches and basic techniques
  - Global Coherence Field (GCF) and Oriented GCF (OGCF)
  - Experimental results and NIST benchmarking
  - Demo
- **Part III: Other features for Acoustic Scene Understanding**
  - Speech/Noise Source Activity Detection
  - Multi-microphone based F0 Estimation + Demo
  - Acoustic Event Classification + Demo
  - Distant-talking Speech Recognition + Demo
  - ASR and Understanding given a Microphone Network

Speech/Noise source Activity Detection (SAD)

- In a real noisy and reverberant environment, SAD is a very challenging task!
- In a real application, a speaker location and tracking system is also characterized by its capabilities to produce in real-time position estimates only when a speaker is active, i.e, reducing false alarms and deletions.
- The peaks of the CSP (or GCF, or OGCF) functions are suitable features in a ***fixed threshold*** -based speech activity detection algorithm [Armani et al. 2003, Brutti et al. 2005].
- In the following example, the speaker was at 3 m distance from the microphones:

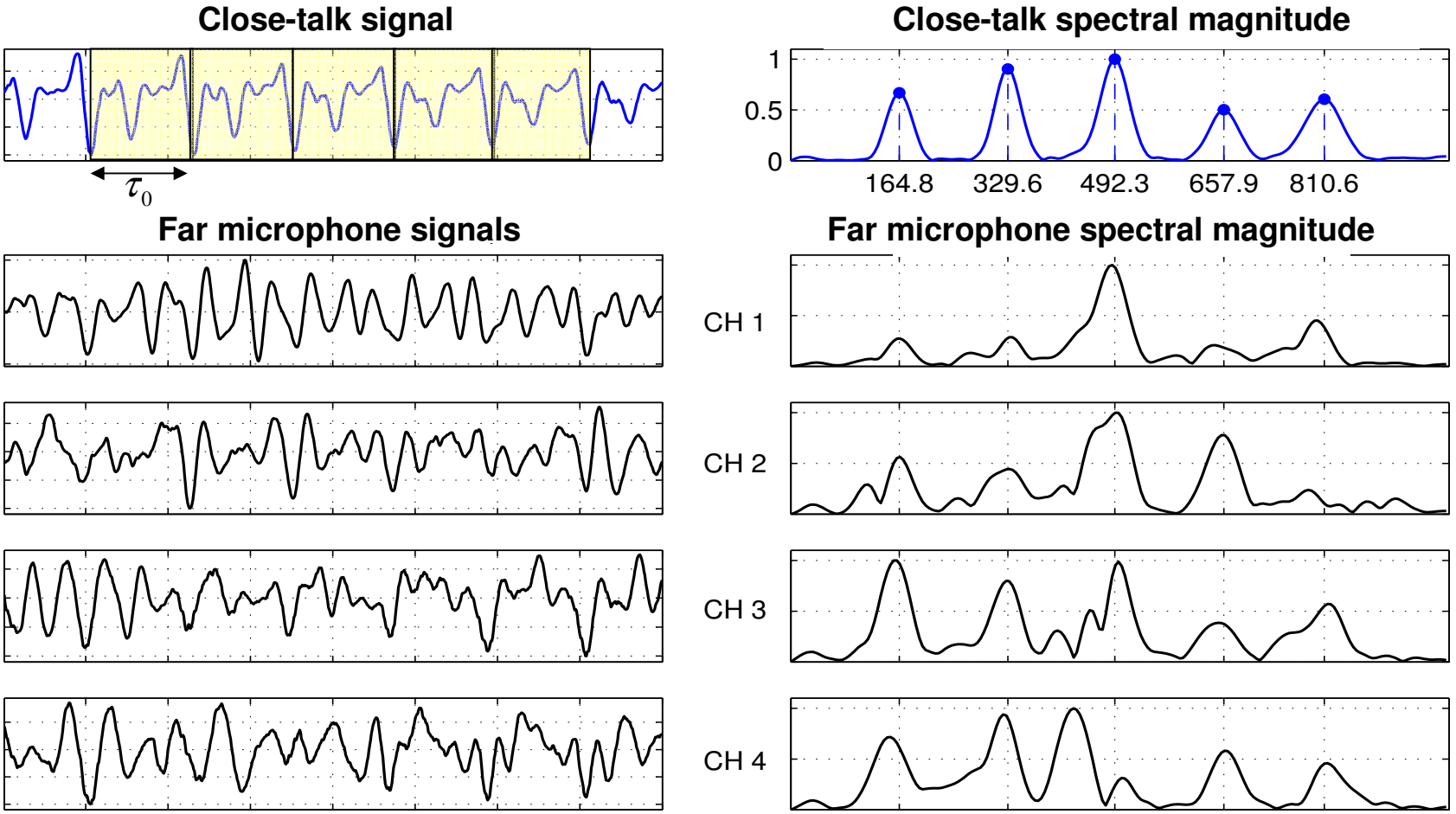


Multi-microphone based F0 estimation algorithms

- **Time-domain processing based algorithms:**
  - **Multi-channel Weighted AUTOCorrelation (WAUTOC),** see [Armani-Omologo, ICASSP 2004] and an introduction to single channel WAUTOC in [Shimamura-Kobayashi, Trans. on SAP, 2001]
  - **Multi-channel YIN,** derived from the single channel YIN algorithm [see De Cheveignè – Kawahara, JASA April 2002]
- **Frequency-domain processing based algorithm**
  - **Multi-channel Periodicity Function (MPF)** algorithm [see Flego-Omologo, EUSIPCO 2006]

Effects of room acoustics

- The periodicity structure in the time-domain, the spectral magnitudes at F0 and at its harmonics can vary a lot!



Multi-channel Periodicity Function (MPF) Algorithm

1) A **windowed version** of the given microphone signals $s_i(n)$ is obtained and denoted as $s_i^w(n)$.
2) A **weighted normalized magnitude spectrum** is computed as: $\bar{S}_{ave}(k) = \sum_{i=1}^{M} \left\{ c_i S_i(k) / \max_l [S_i(l)] \right\}$

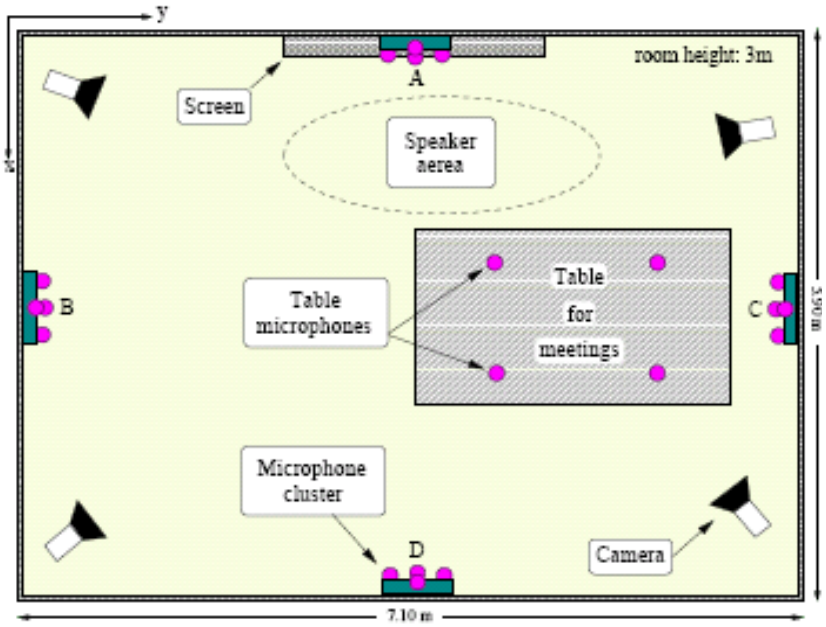
   given: $S_i(k) = \left| FFT\{ s_i^w \}(k) \right|$

   where the generic weight c_i denotes the *reliability* of the related *i*-th channel.
3) An **inverse FFT** is computed from the average spectrum: $\bar{s}(\tau) = IFFT\{ \bar{S}_{ave} \}$

   $$\hat{\tau}_0 = \arg\max_{\tau} \{ \bar{s}(\tau) \}$$
4) The resulting function is **maximized**, i.e.

Experimental framework for multi-microphone F0 estimation

1) Keele database reproduced by a loudspeaker in an office environment
2) Simulations
3) **CHIL real seminar corpora**:
   - 16 omnidir. microphones
   - ~7m x ~6m x 3m,T60~0.4s
   - Reliable F0 estimate available as reference

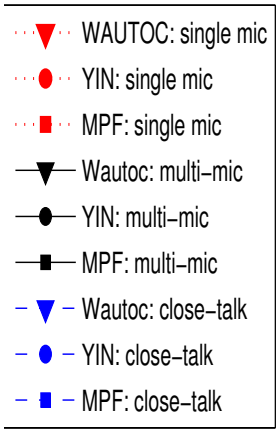


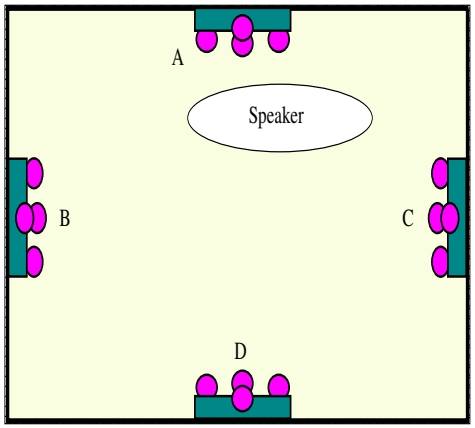
Evaluation criteria: Gross Error Rate (GER) given a percentage of tolerance

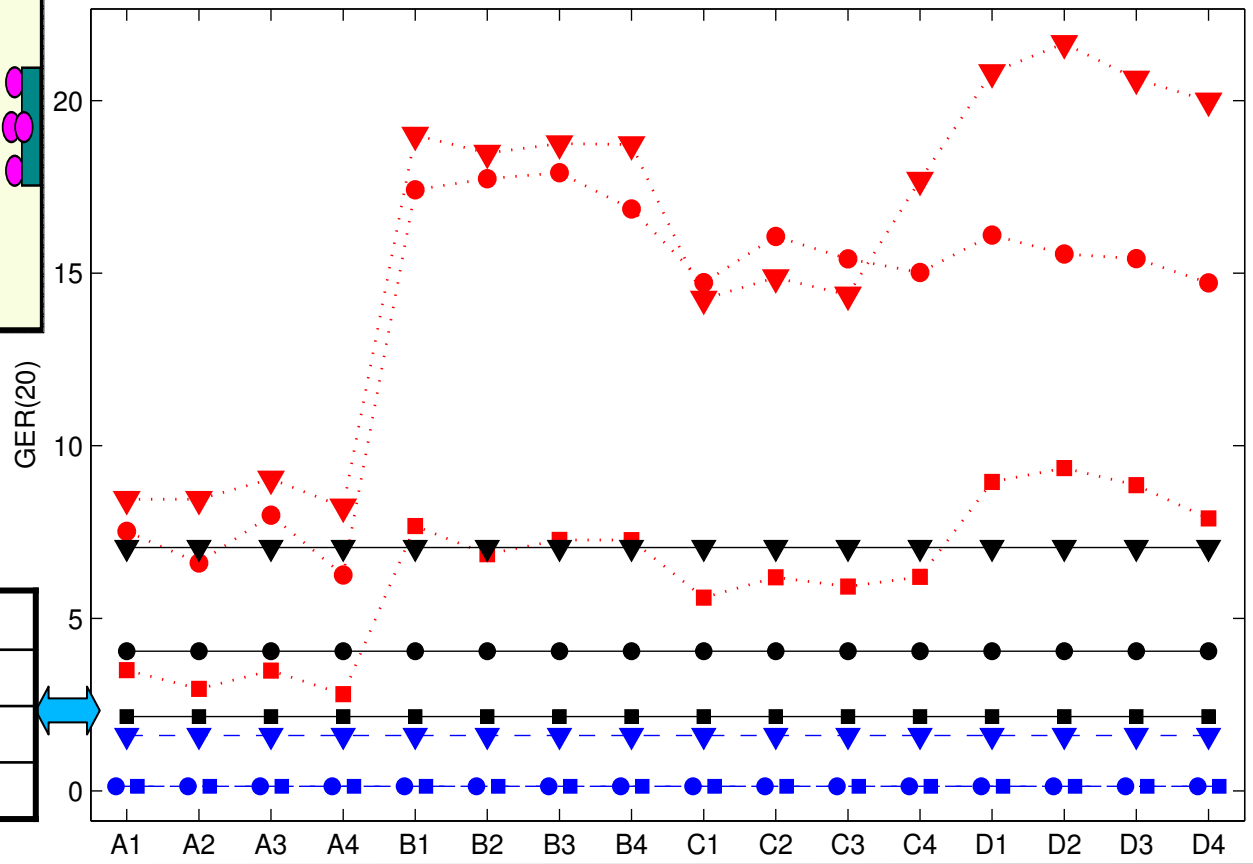
$$GER(\theta) = \frac{100}{N_{fr}} \sum_{i=1}^{N_{fr}} \left\{ \frac{\left| \hat{F}_0(i) - F_0(i) \right|}{F_0(i)} > \theta\% \right\}$$

Comparison between MPF, WAUTOC, and YIN

- Gross Error Rate (20%) evaluated on the CHIL corpus: close-talk vs single far microphone vs multi-microphone







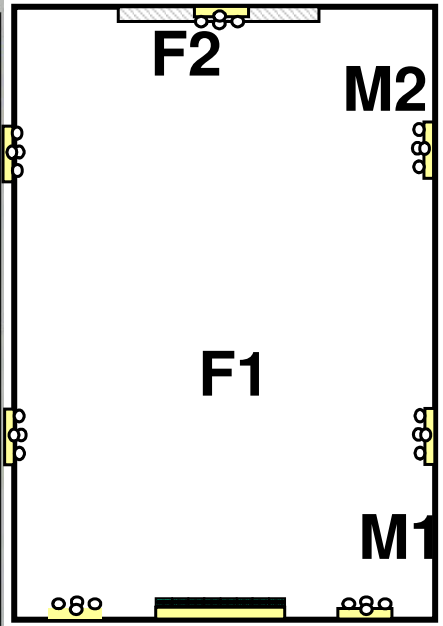
GER(20%)	WAUTOC	YIN	MPF
CloseTalk	1.60	0.13	0.14
SingleMic	15.84	13.83	6.30
MultiMic	7.05	4.05	2.15

Demo 2: Multi-microphone based F0 Estimation

Real-time F0 estimation given a range between 50 and 400 Hz

This video-clip can be downloaded from http://shine.itc.it

Map of the CHIL room at ITC-irst



- 2 male and 2 female (2 italian and 2 english) speakers
- 2 repetitions of the “aeiou” sequence
- MPF based on 6 microphone pairs
- Real-time F0 tracking

Acoustic Event Detection and Classification in CHIL

- Focus of CHIL on: seminar and meeting scenarios
- ***Isolated*** and ***connected*** event recognition tasks:
  - Vocabulary of 12 semantic classes (e.g., speech, knock, paper wrapping, applause, cough, phone ringing, etc.)
  - Scripted recordings of at least 50 occurrences/event (in isolated mode) with Distributed Microphone Networks available at UPC and ITC-irst
  - Additional data recorded in real interactive seminars (immersed in the real speech of other sessions at UPC)
  - Manual labeling and evaluation metrics definition
- AED/C CLEAR evaluation carried out in February 2006 (3 participants: UPC,CMU, and ITC-irst)

Acoustic Event Detection and Classification at ITC-irst

- Investigated technique:
  - One HMM per event (+silence and speech)
  - HMM topology: 3 states, each with one Gaussian
  - Acoustic features: 12 mel cepstral coefficients + energy + first and second order derivatives
- Preliminary experimental results (event class.error rate):

System / Corpus	ITC-isol.	ITC-con.
ITC	12.3%	**23.6%**
UPC	6.2%	**33.7%**

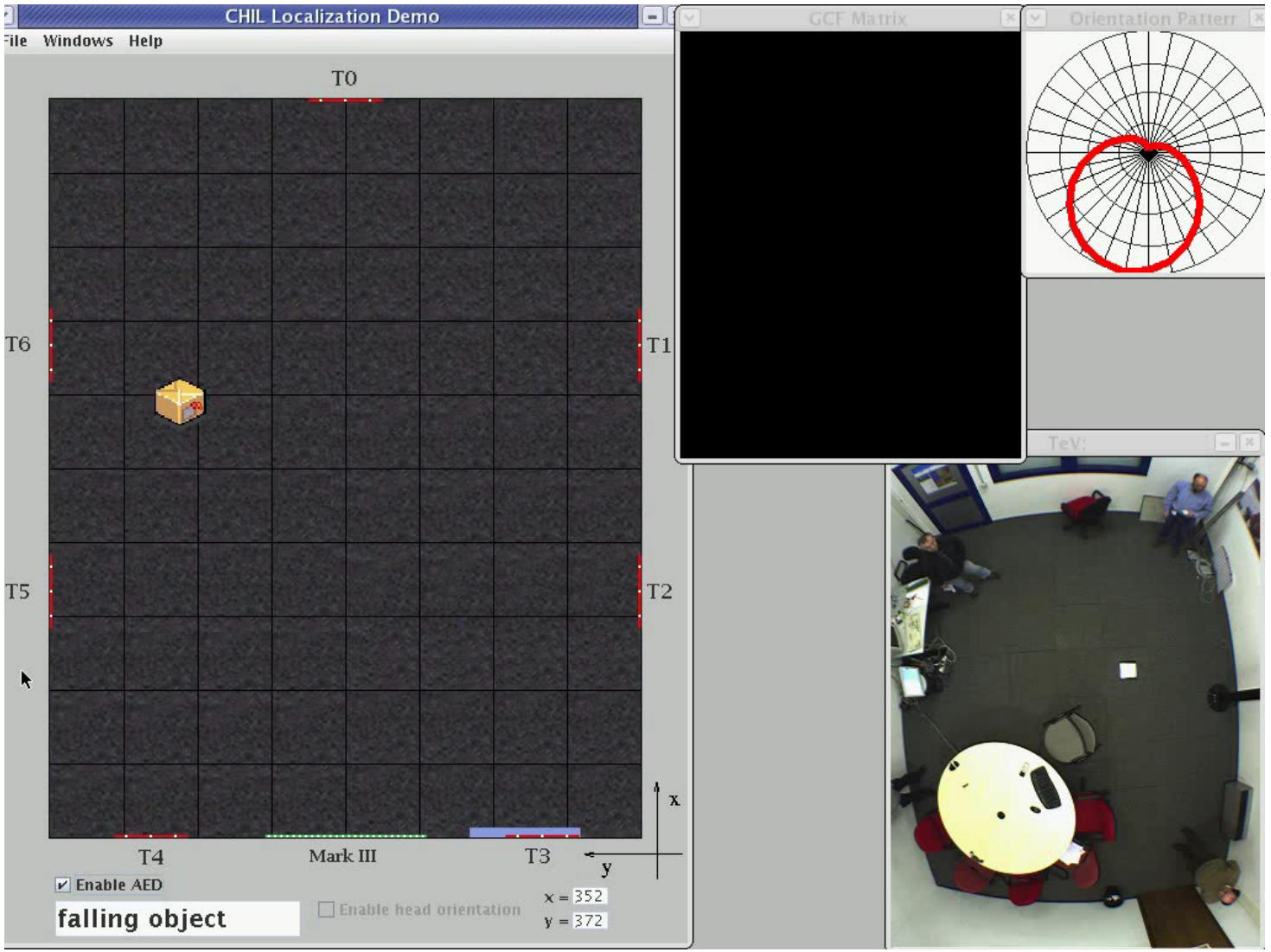
- Error rates in real interactive seminars close to 100%!...
- Very difficult task: need of investigating on other methods and, in particular, on other acoustic features

Acoustic Event Detection and Classification at ITC-irst

- Other features being investigated:
  - Estimation of the number of active speech/noise sources
  - Time structure of the event (e.g. contours of energy and other features)
  - Periodic vs aperiodic characteristics (e.g. MPF and related information)
  - Estimated radiation information (e.g. OGCF)
  - Magnitude Squared Coherence
- Other pattern matching techniques:
  - Machine learning, Support Vector Machines, etc
- Other tasks:
  - New real corpus collected in the CHIL room at ITC-irst, which contains an extended set of events (e.g. sound and noise produced by a MIDI expander and diffused through a louspeaker)

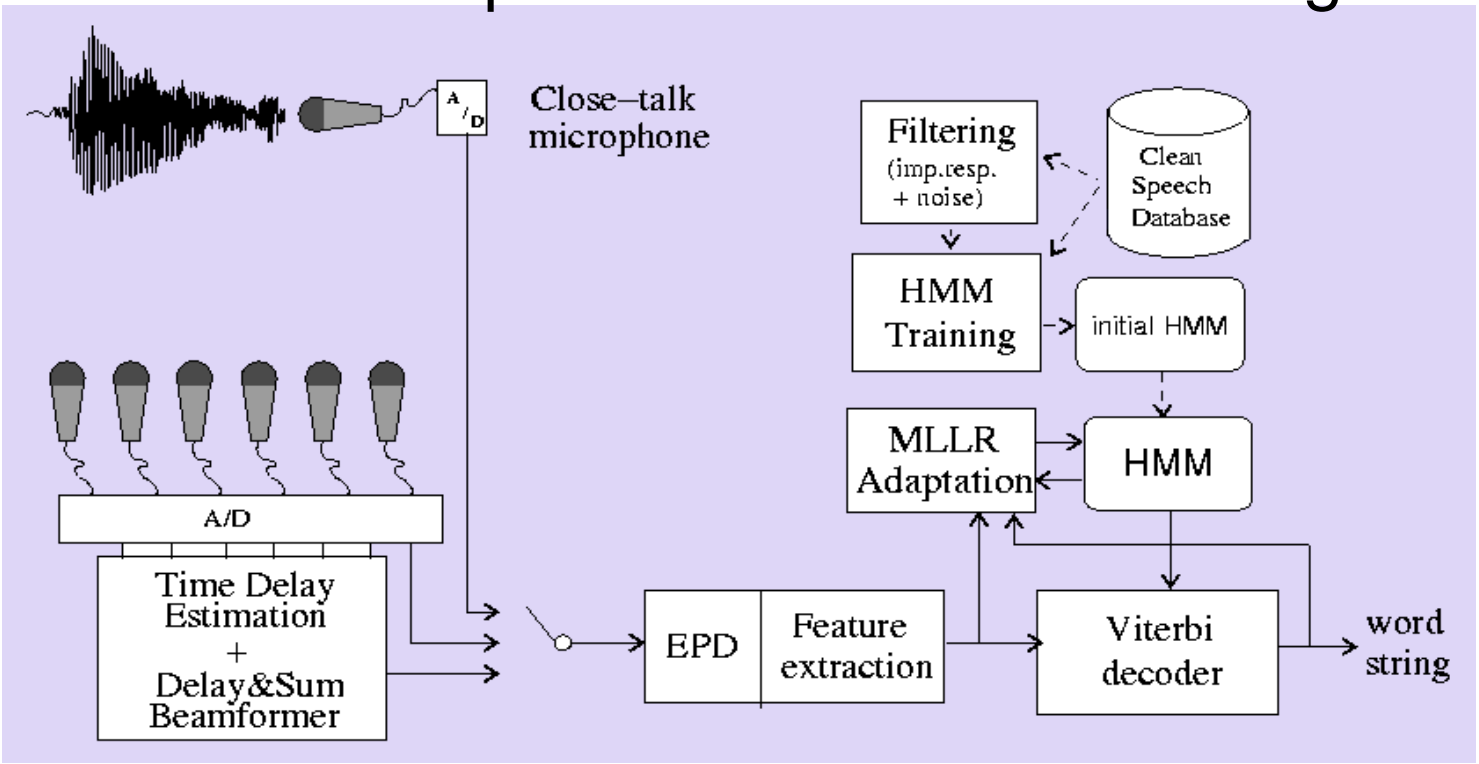
Demo 3: Acoustic Source Location and Event Classification

- Combination of a SLOC system and an acoustic event detection and classification system
- Event classification based on HMMs (given a vocabulary of 15 events)

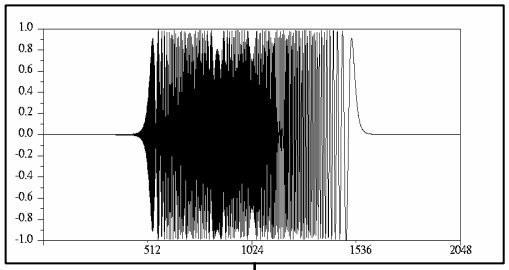


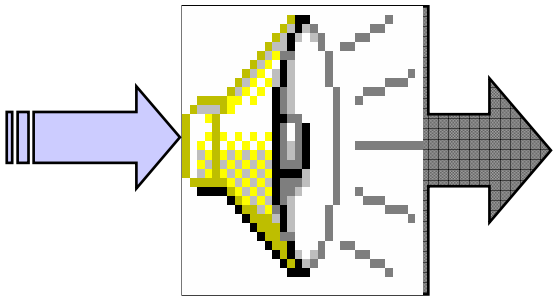
Distant-talking ASR

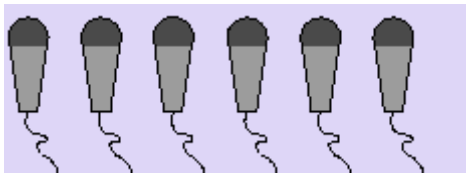
- Previous work at ITC-irst based on:
  - a linear microphone array
  - GCC-PHAT based D&S beamforming
  - contaminated speech based HMM training



Training of HMMs with contaminated speech



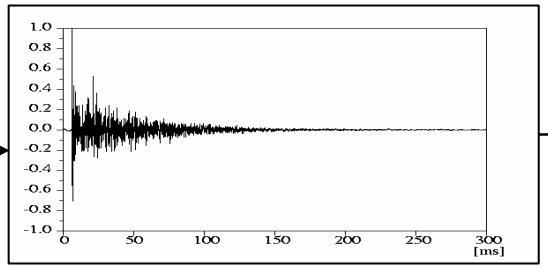




Chirp-like signal

A/D

Cross-correlation



Impulse response

Clean speech

Convolution

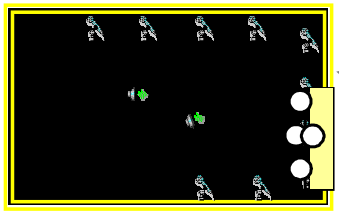
Sum

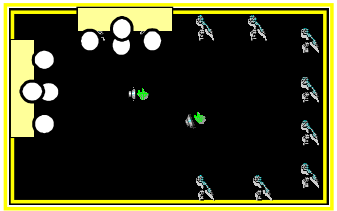
Env. Noise at various SNRs

HMM training

HMMs

Speech recognition and understanding given a microphone network





Multi-microphone parallel recognizers

ASR 1

ASR 2

ASR N

AM/LM

AM/LM

AM/LM

Preprocessing (SAD, OGCF, etc), dynamic selection of microphone clusters, ac. feature extraction

Microphone Signals

Information from subsets of microphones

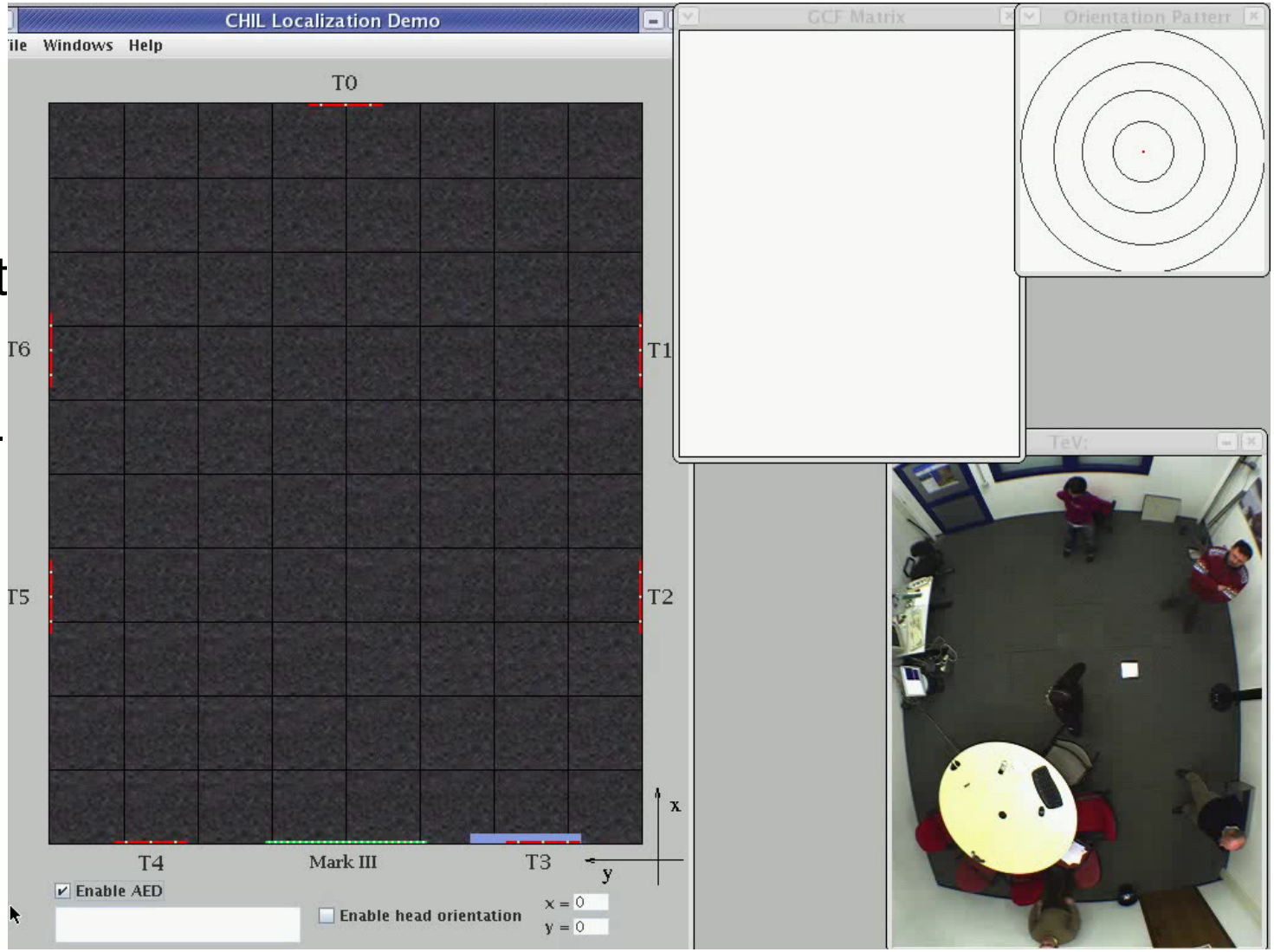
Selection, Fusion, Discrimination based on confidence measures

Recognized sentences

N-best sentences or Word Hypothesis Graphs

Demo 4: SLOC and Distant-talking Speech Recognition

- GCF-based speaker location
- 4 speakers
- Connected digit recognition task
- Current distant-talking recognition performance in the CHIL room at ITC-irst: ~90% digit accuracy under real interaction



Maurizio Omologo – omologo@itc.it

Future work: relevant research areas towards ambient intelligence

- Audio-video sensor processing integration (e.g. OGCF and visual maps) ➡ sinergy between independent information
- Integration with other sensor technologies, multimodality
- Acoustic features for distant-talking ASR and event classification
- Large vocabulary distant-talking speech understanding and dialogue
- Speaker variability in applications of distant-talking ASR
- Parallel recognizers
- Combining Speaker Location and Speaker Identification
- Address overlapping speakers in cocktail party scenarios
- Detection and Classification of acoustic events *immersed* in noise and speech ➡ very challenging task

THANK YOU

FOR YOUR ATTENTION!

Maurizio Omologo – omologo@itc.it