Multi-microphone signal processing for distant-speech interaction

Maurizio Omologo

Fondazione Bruno Kessler - irst, Trento, Italy

Human Activity and Vision Summer School (HAVSS)

INRIA Sophia Antipolis, October 3rd, 2012

Outline

- Introduction
- DIRHA EC project
- Some basic methods of signal processing
- Sound propagation in an enclosure
- Single speaker localization and tracking
- Estimation of the head orientation
- Multiple speaker localization and tracking
- Source separation and extraction
- Distant-speech interaction: DICIT EC project
- Demo video-clips

Introduction

- Speech: the most accessible, natural, *easy-to-use* interface
- Attractiveness and usefulness of distant-talking automatic speech recognition (ASR) interfaces
- From close-talking to distant-talking ASR: a very challenging task
- Complexity of the problem due to environmental noise, room acoustics, interfering speakers, etc.
- Need to “understand” the acoustic scene in real-time before applying ASR
- Large number of possible applications

The DIRHA project

- DIRHA project - Collaborative Project – STREP
- FP7- ICT - 2011 - 7 - Language technologies
- Duration: 36 months
- Start date: 1 January 2012
- Consortium composition:

FONDAZIONE BRUNO KESSLER
Trento, Italy

ATHENA RC-IAMU
Athens, Greece

inesc id lisboa
Lisbon, Portugal

new amuser advanced multimedia services
Torino, Italy

TU Graz
Graz, Austria

DomoticArea S.r.l. HOME & BUILDING AUTOMATION
Rovereto, Italy

Milano, Italy

For more details see the web site http://dirha.fbk.eu

General goals of the project – at scientific-technological level

Acoustic scene analysis | **Distant-speech interaction**

↓

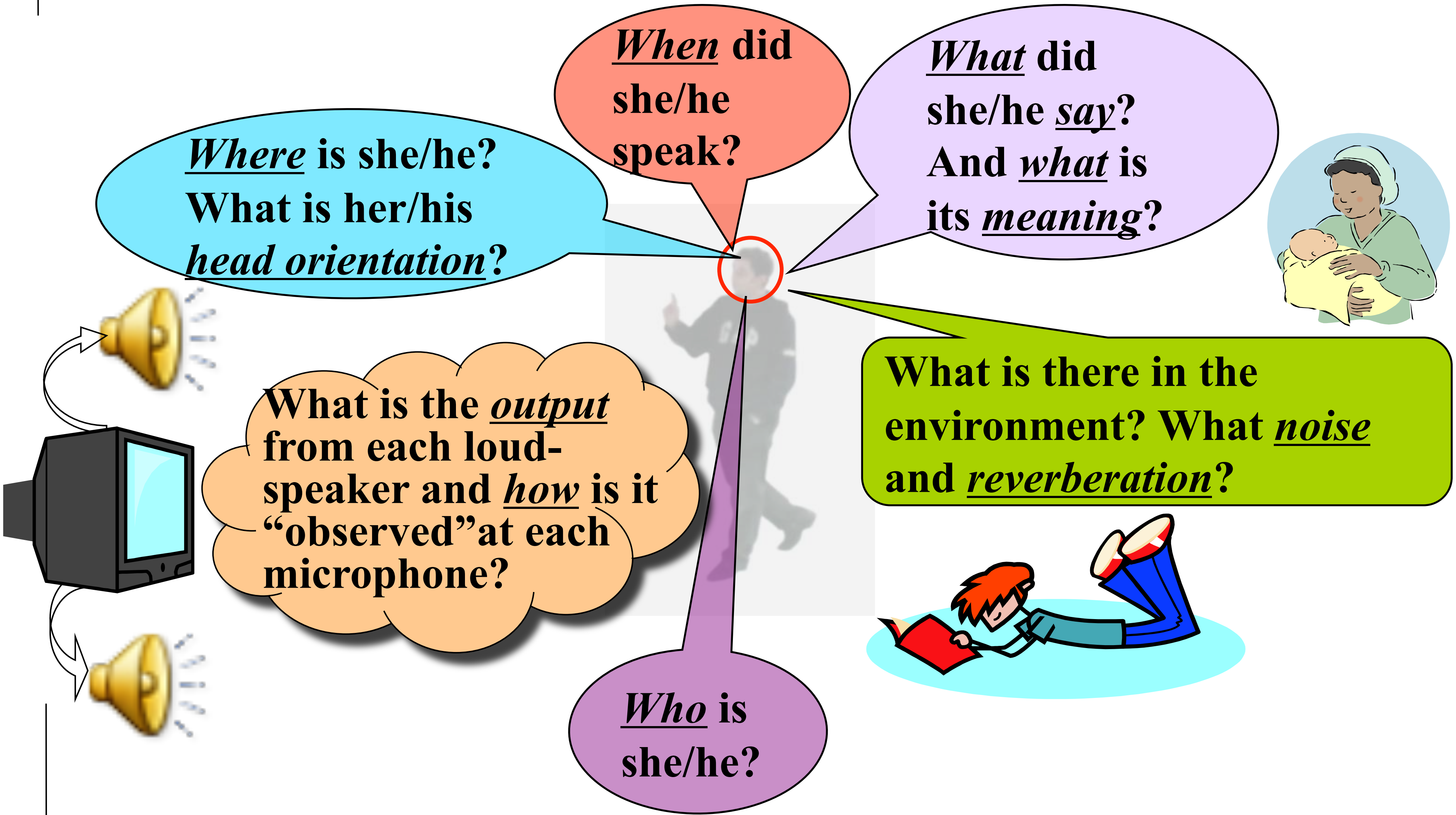
Voice-enabled home automation

- Distributed microphone network
- Always listening system
- No-push-to-talk activation
- Multi-room multi-speaker/sound sources
- Robustness to any conditions of a domestic context
- Multi-language system [En, Ge (Au-D), It, Gr, Pt]
- Motor-impaired end-users
- Real installation in end-user homes
- Full integration with home automation systems

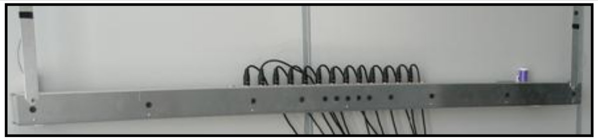
Possible tasks and scenarios

- Control doors, lights, shutters, air-conditioning, temperature
- Emergency, alarm management
- Phone calls, entry-phone and other communication means
- Control of radio, TV, HiFi, PC, etc.

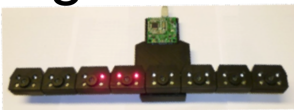
Acoustic scene analysis (ASA): basic functionalities to realize



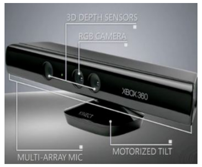
DIRHA lab: ITEA apartment



- Microphone arrays and microphone pairs on walls and ceilings
- MEMS digital microphone arrays

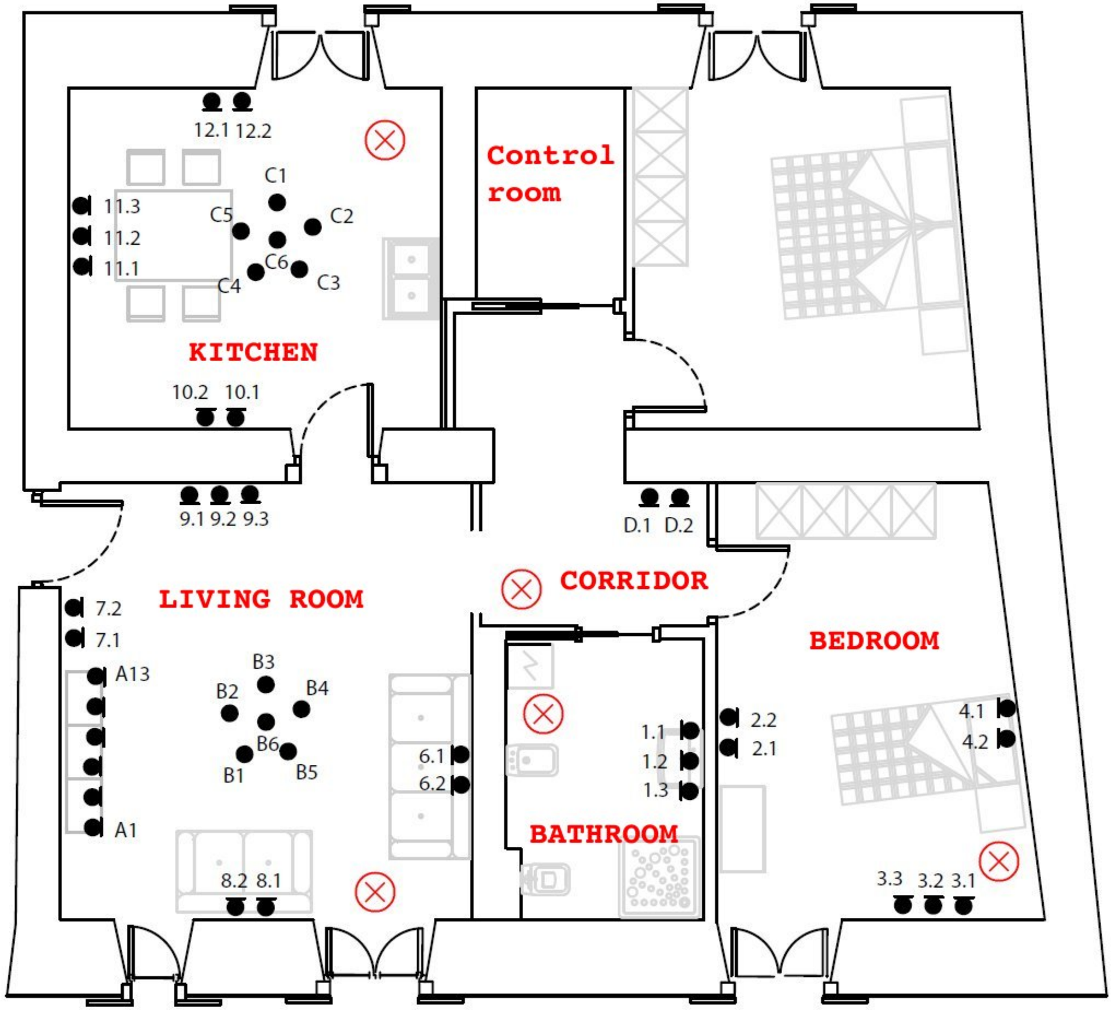


- Kinect devices



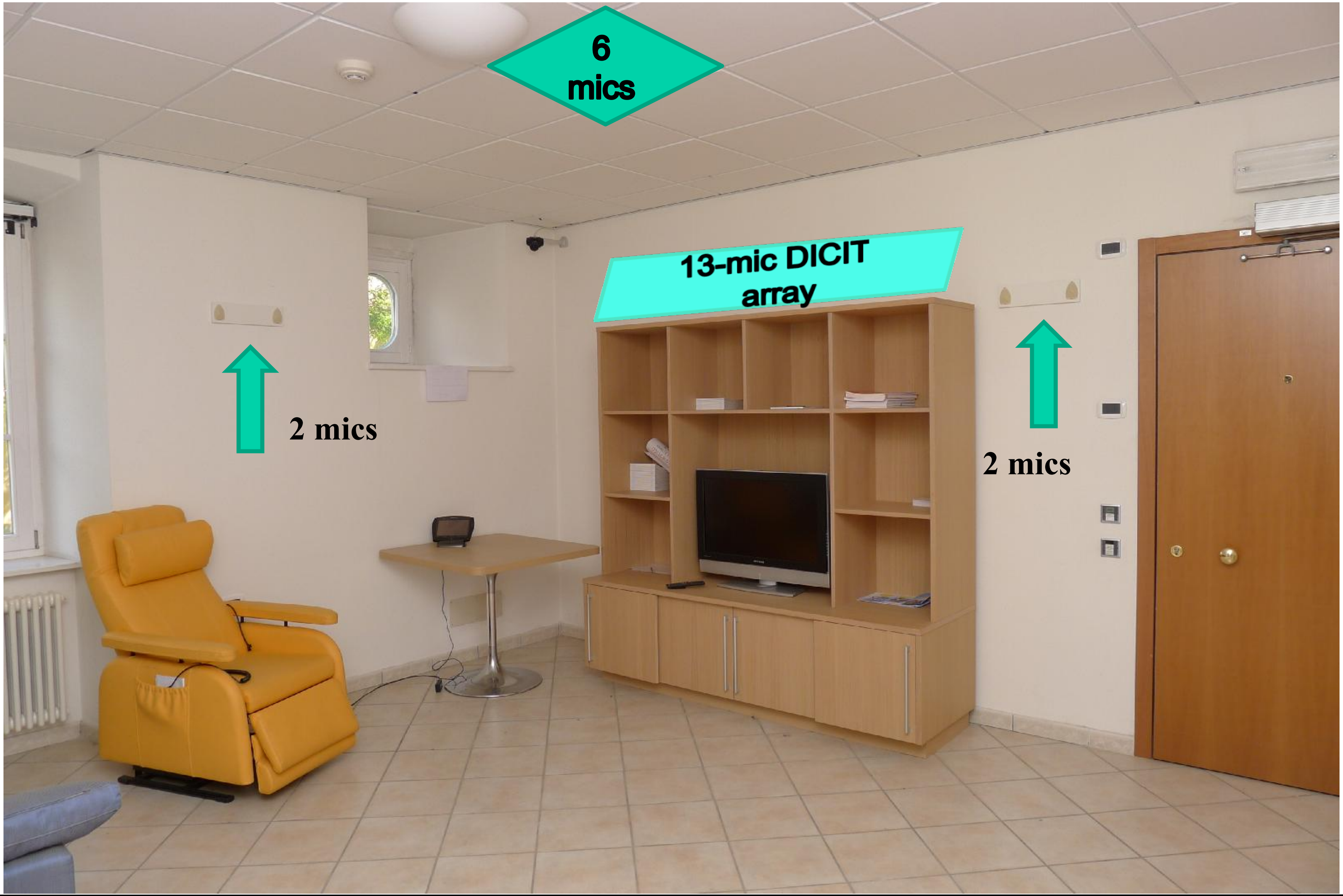
- Loudspeakers (one per room)
- Intra-phone, entry-phone, TV, HiFi, etc.

Most of them integrated in the same audio acquisition framework (Fs=48kHz, A/D at 24bit)



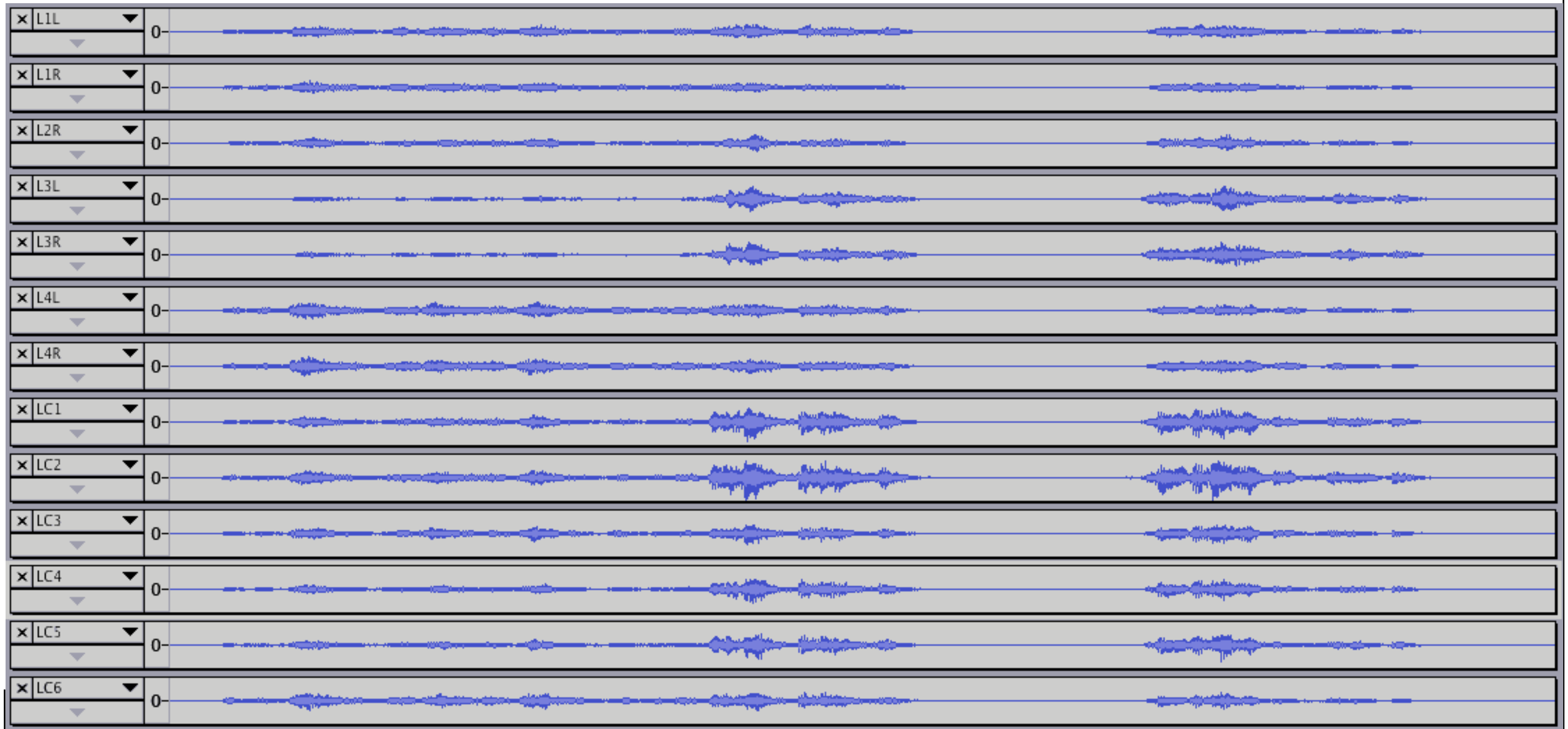
Maurizio Omologo – omologo@fbk.eu

ITEA apartment – living room



Example of multi-channel speech sequence

Speech signals recorded in the living room

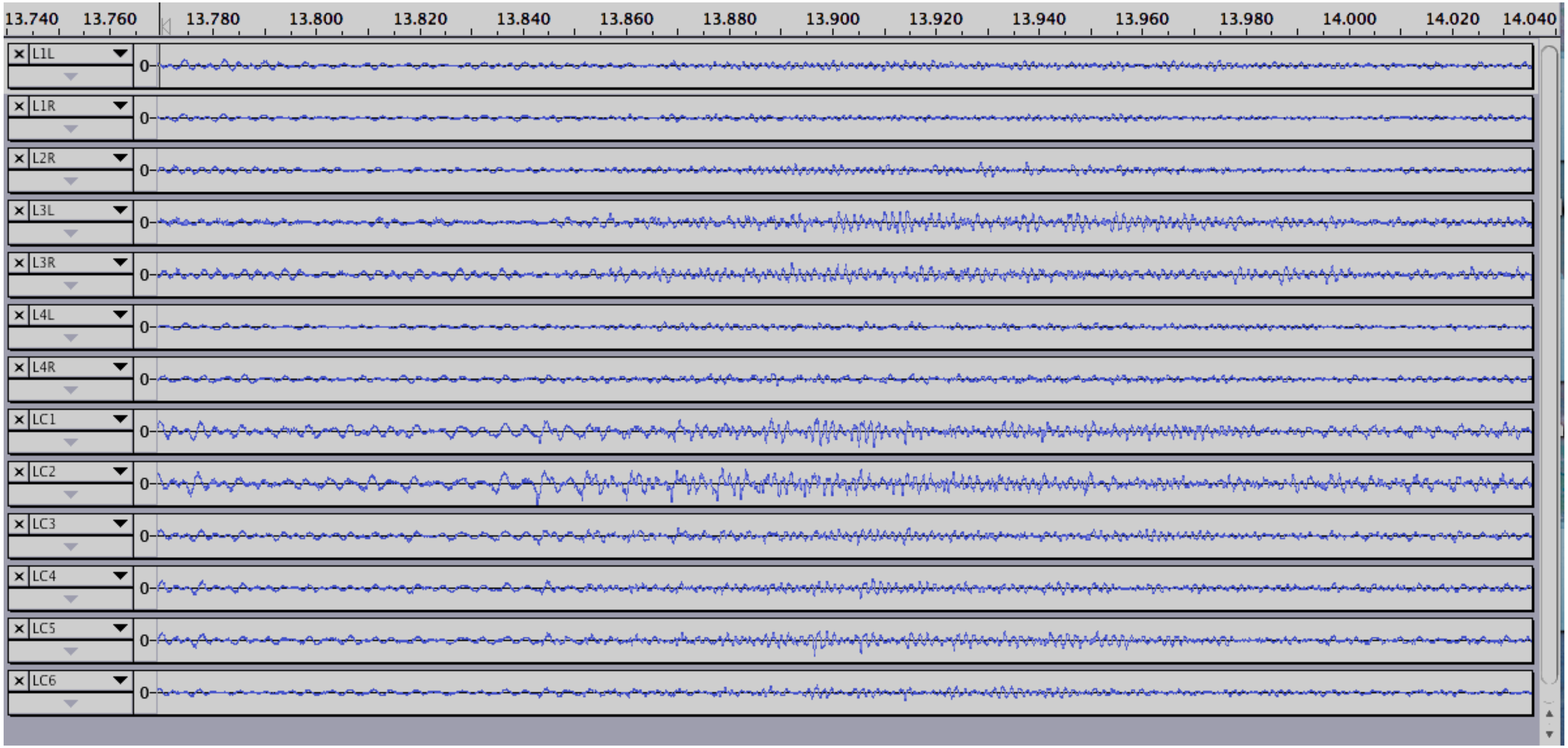


Discrepancies in dynamics between different channels

Maurizio Omologo – omologo@fbk.eu

Example of multi-channel speech sequence

Zoom of a short segment of about 300 ms



- It's clear that the signals are very different one to another

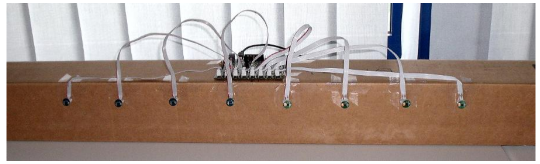
How to process all these signals?

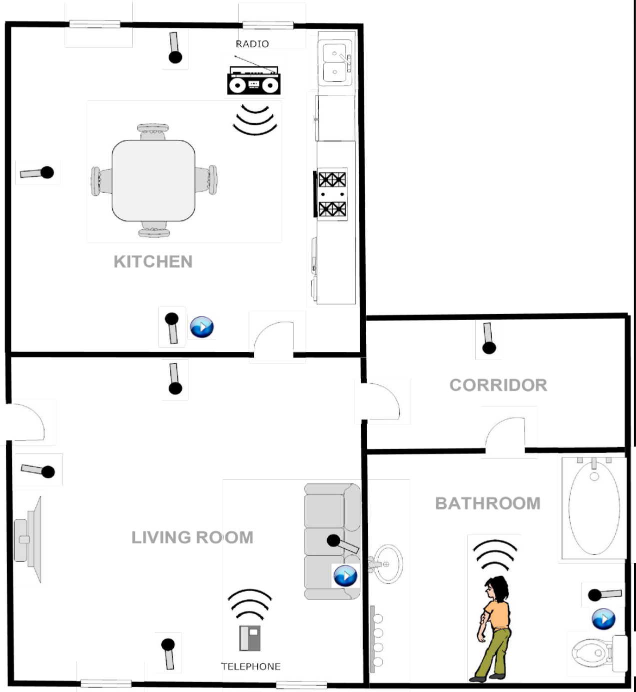
- Differences among signals: the scene is the same, but sampled at different observation points
- Complexity of the scene: in general, mix of different source activities
- Goal: extract a coherent analysis from this network of sensors
- Approach:
  - Low level signal processing for microphones close each other
  - Higher level processing for microphone clusters far each other
  - Integrate multi-channel signal processing with techniques for classification, recognition, understanding, etc.
- Constraints:
  - Limited number of microphones
  - Real-time processing for prototype development
  - Synchronized platforms (using the same clock)

Main scientific and technological fields

ASA

- Microphone array processing, use of MEMS digital microphones
- Sound source localization and tracking
- Source separation and enhancement
- Multi-channel acoustic echo cancellation
- Acoustic event detection-classification
- Distant-speaker ID/verification
- Distant-speech recognition, keyword spotting
- Spontaneous natural speech understanding
- Concurrent dialogue management
- Response generation, feedback to the user





Robustness of all the involved technologies

Past EC projects - DICIT, CHIL, and SCENIC

DICIT (Distant-talking Interfaces for Control of Interactive TV) http://dicit.fbk.eu

- STREP Project – FP6 - 2.5.7 Multimodal Interfaces
- Duration: October 2006 – September 2009
- Coordinator: FBK (I)
- Other partners: Amuser (I), Elektrobit (D), FAU (D) Fracarro (I), IBM (CZ, USA)

CHIL (Computers in the Human Interaction Loop)

- Integrated Project – FP6 - IST-2002-2.3.1.6
- Duration: January 2004 – August 2007
- Coordinator: Karlsruhe University (D)
- Consortium consisting of 17 partners

SCENIC (Self-Configuring ENvironment-aware Intelligent aCoustic sensing)

http://www.thescenicproject.eu

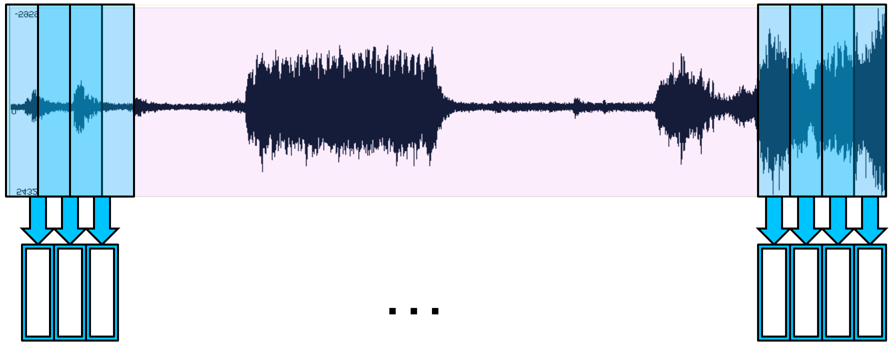
- FET Open - STREP – FP7
- Duration: January 2009 – December 2011
- Coordinator: Politecnico di Milano (I)
- Other partners: Imperial College of London (UK), Fondazione Bruno Kessler-irst (I), Friedrich-Alexander Universtaet Erlangen-Nuernberg (D)

Some basic audio signal processing methods

Audio and speech signal processing: basic notions

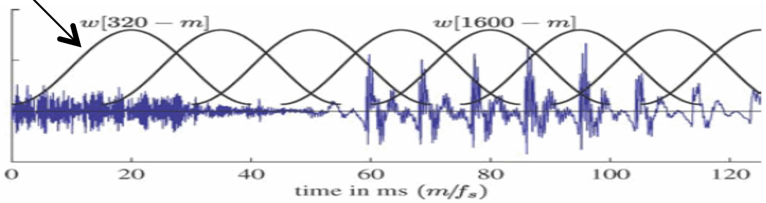
- Changing speaker position and/or head orientation the microphone signal differs substantially
- Differences are significant, even when a speaker repeats the same utterance standing at the same position!
- Non-stationarity of most of the processes generating audio activities embedded in the acoustic scene
- Need to do an assumption of local stationarity (or quasi-stationarity) and analyse short intervals
- Analysis step and window size from 10 to 200-300 ms according to the problem

Most common approach:
Short-time Fourier analysis from which one derives a sequence of feature vectors



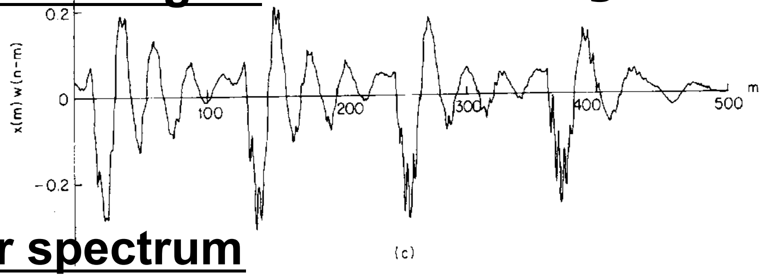
Short-time Fourier analysis

- Useful to deal with time-varying properties of signals
- Based on windowing a given temporal interval of $x(n)$
- Defined as follows: $$X_n\left(e^{j\omega}\right) = \sum_{m=-\infty}^{\infty} w(n-m)\,x(m)\,e^{-j\omega m}$$

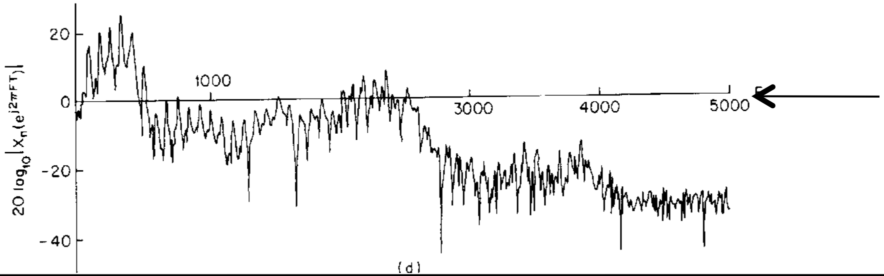


- Effect of windowing on power spectrum

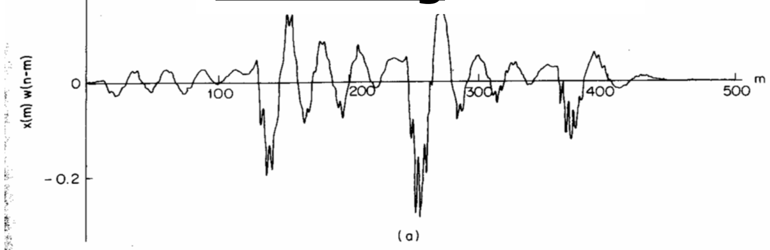
Windowed signal **Rectangular**

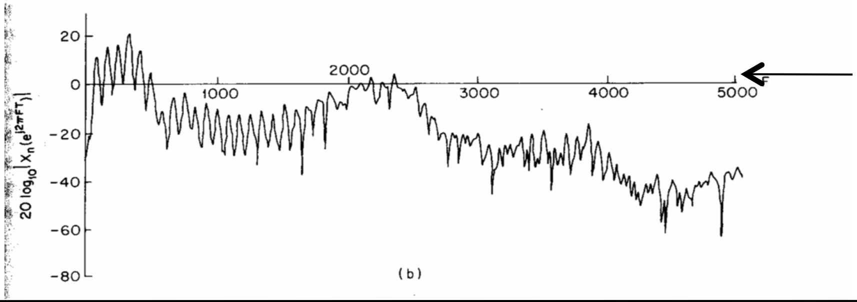


Power spectrum



Hamming

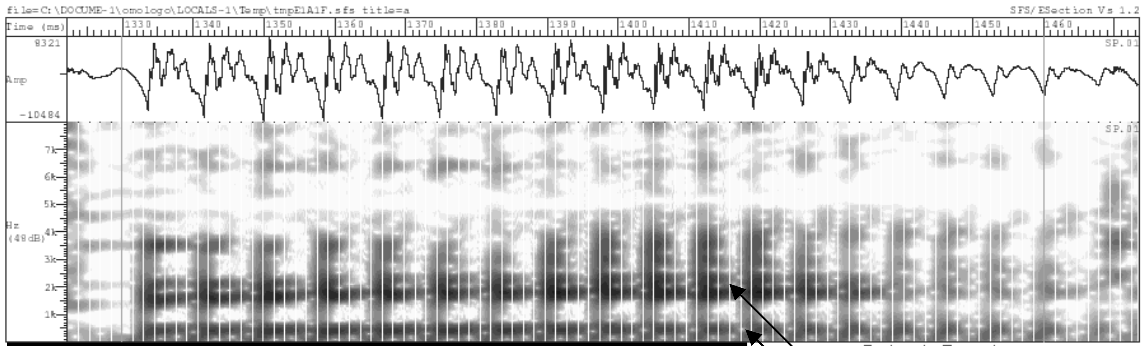


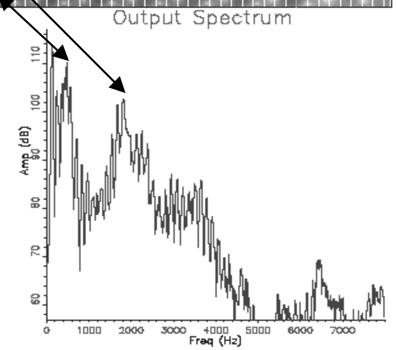


Spectrogram: a very common tool for speech analysis

Since the 1940s, the ***Spectrogram*** has been a basic tool for gaining understan-ding of how speech is produced and how phonetic information is encoded in it

- It consists in a gray-scale or a color-mapped image on a time (x axis)-frequency (y axis) plane.
- The gray or color intensity denotes the magnitude of the Short-Time Fourier Transform of the given signal segment for a given time instant and frequency

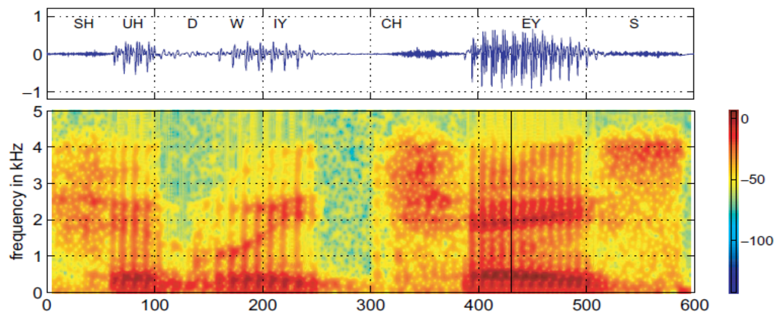




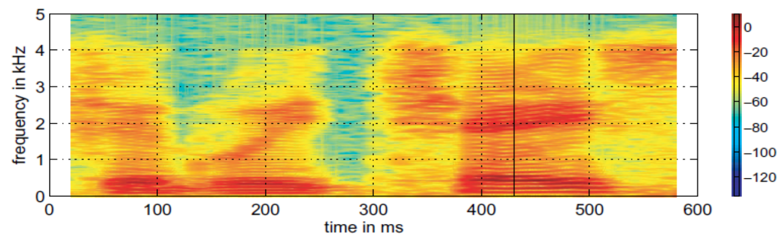
<u>Wideband</u>: STFT analysis done with a short window (e.g. 15 ms) – i.e. a broad bandwidth of the filtering in frequency

<u>Narrowband</u>: larger window size (e.g. 50 ms)

- Examples of **wide-band spectrogram**



… and **narrow-band** one

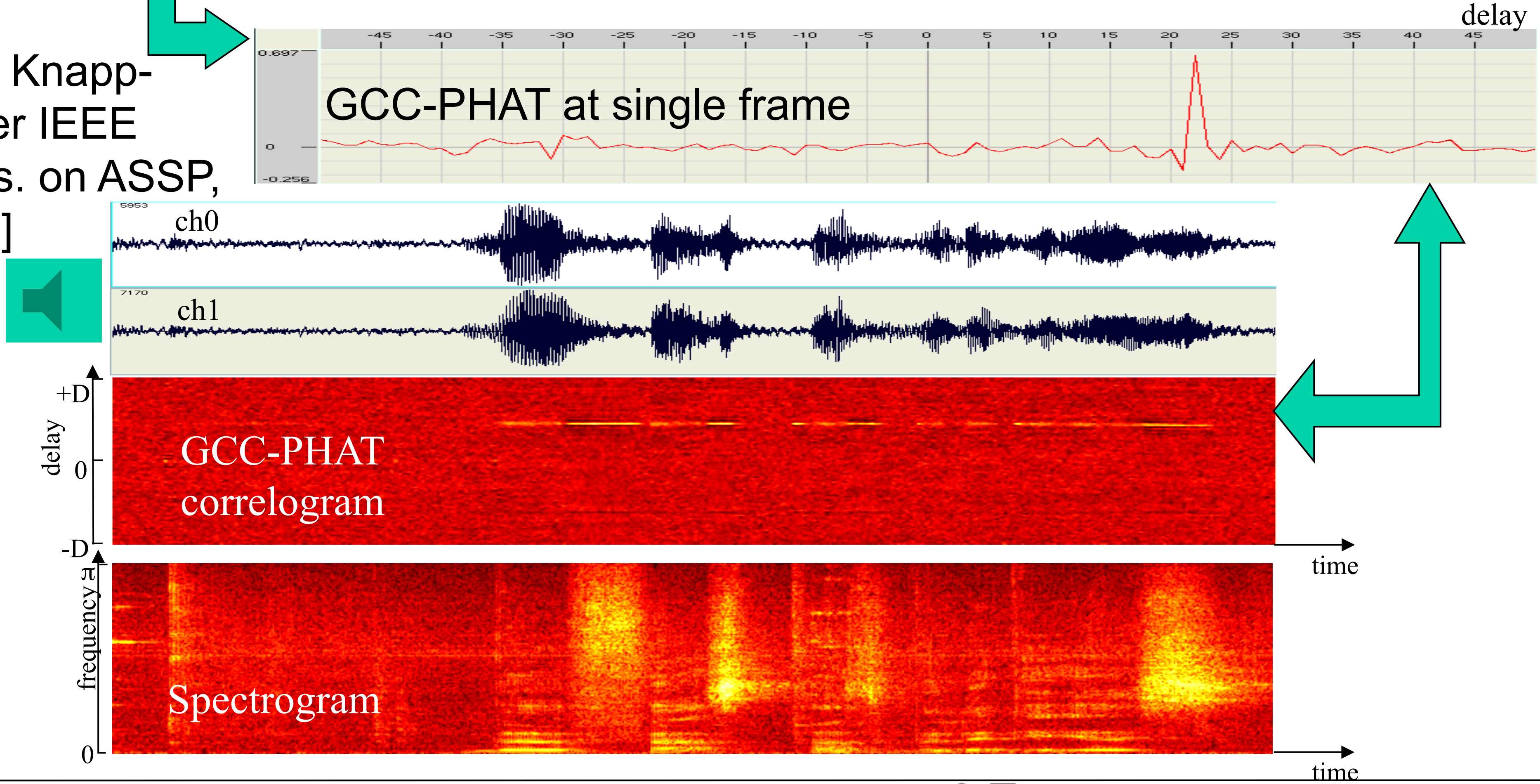


Given two microphone signals: GCC-PHAT analysis

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

where $Y_0(\omega), Y_1(\omega)$ denote the short-time Fourier transform of the two microphone signals for the current frame

[see: Knapp-Carter IEEE Trans. on ASSP, 1976]



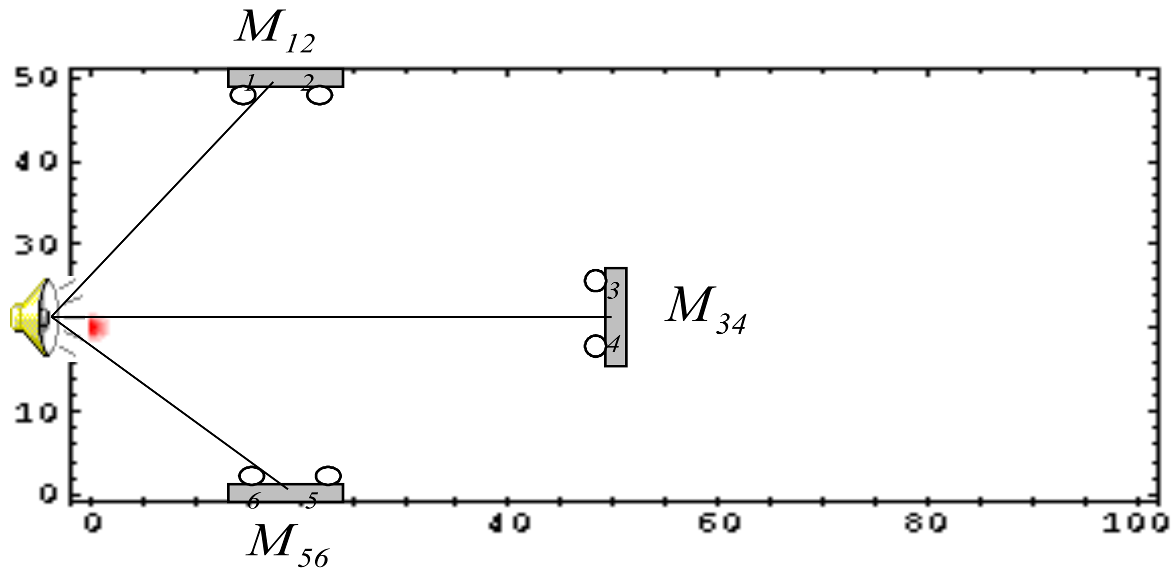
TDOAs at microphone pairs

d = the distance between the microphones of a pair = 12 cm

c = speed of sound = 340 m/s

$$t_1 = 0.97\ ms \qquad t_2 = 1.16\ ms$$
$$t_3 = 1.48\ ms \qquad t_4 = 1.48\ ms$$
$$t_5 = 0.96\ ms \qquad t_6 = 0.71\ ms$$

$$\delta_{12} = 0.19\ ms$$
$$\delta_{34} = 0\ ms$$
$$\delta_{56} = -0.25\ ms$$

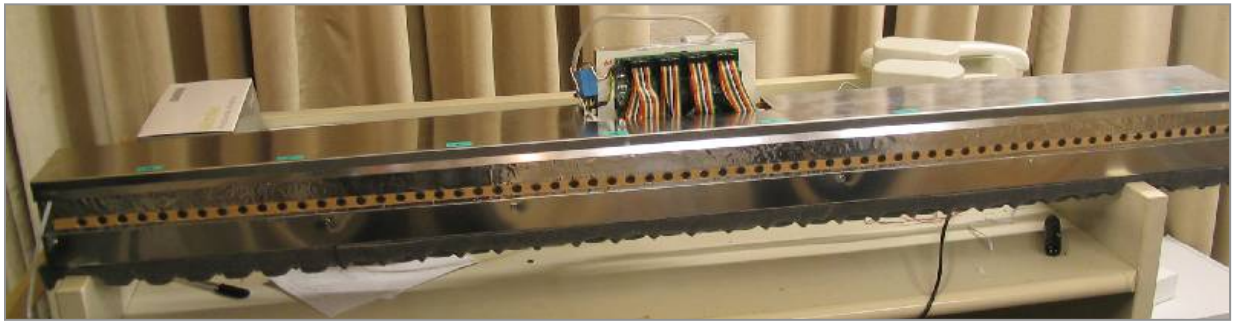


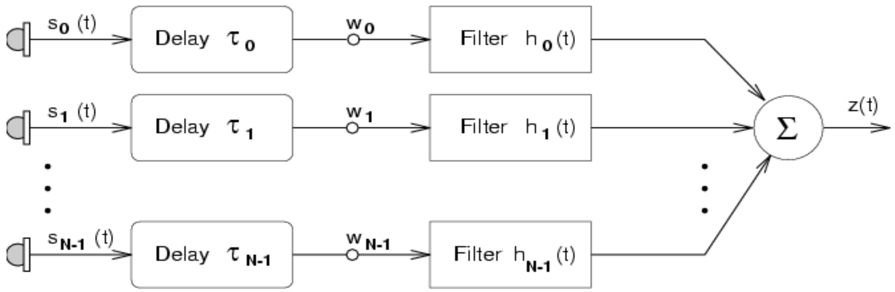
TDOA at each microphone of a pair can be computed on the basis of coherence in direct wavefront highlighted by a peak of GCC-PHAT

Animation courtesy of Dr. Dan Russell, Kettering University

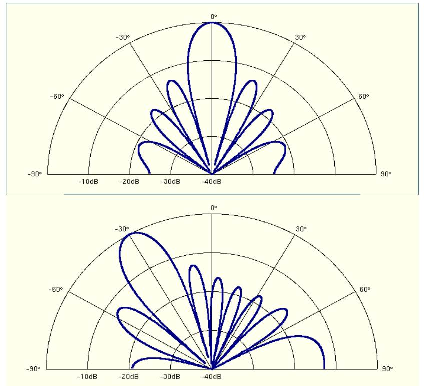
Microphone array processing

- Microphone arrays are multichannel acquisition devices that allow sampling an acoustic field:
  - in time (synchronously)
  - in space (with proper geometry)





- Spatial/temporal filtering allows one to:
  - change the directivity of sound acquisition
  - selectively pick-up and enhance the desired signal
  - cancel or attenuate undesired disturbances
- Solid theory available in the literature, many methods successfully applied (e.g. speech enhancement)



Sound propagation in an enclosure

Acoustic signal modeling

Source at $\mathbf{s} = [s_x, s_y, s_z]$))) r_i ⟶

$x(t)$ = source signal

$y_i(t)$ = microphone signal

Microphone M_i at $\mathbf{m}_i = [m_{ix}, m_{iy}, m_{iz}]$

Speed of sound:
$$c = 331.45\sqrt{\frac{T}{273}}\,(m/s)$$

- Considering attenuation and delay of propagation in a free-field anechoic condition, the simplest model is:

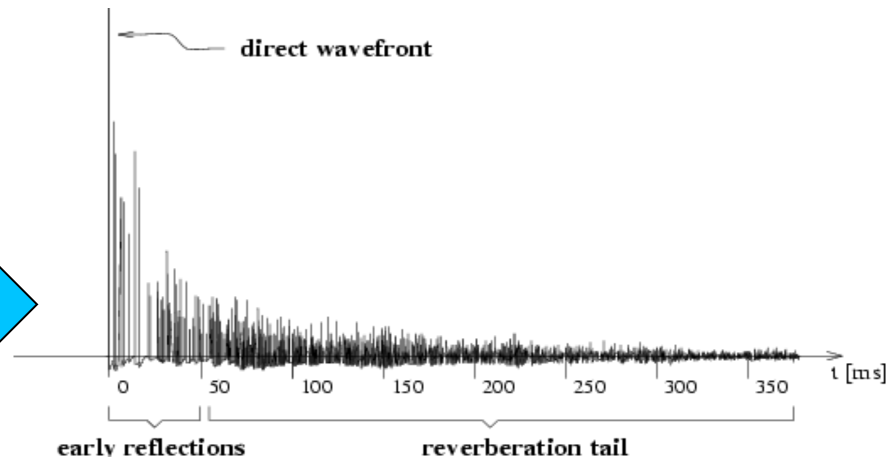
$$y_i(t) = \frac{K}{r_i}x(t - T_i)$$

$$T_i = \frac{|\mathbf{s} - \mathbf{m}_i|}{c} = \frac{r_i}{c}$$ = propagation time (i.e., time of flight)

- In a real, noisy and reverberant environment, taking into account the multiple paths due to sound reflections on surfaces, a more realistic model is:

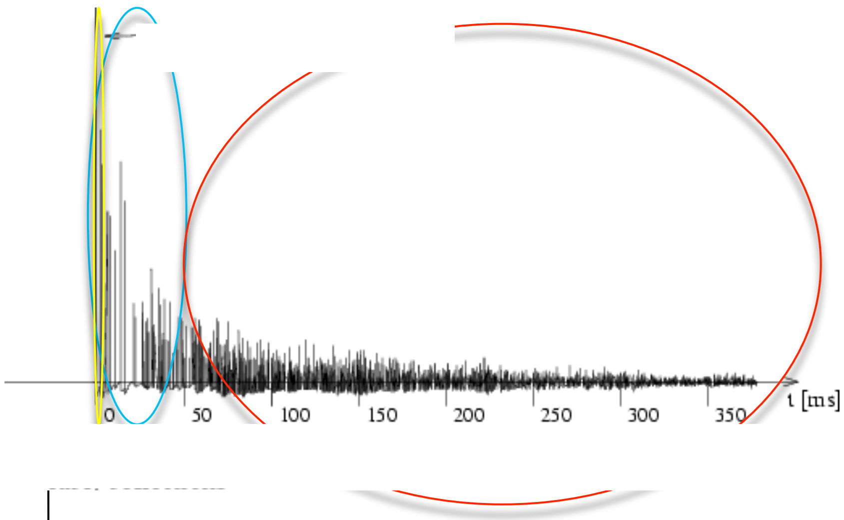
$$y_i(t) = x(t) * h_i(\mathbf{s},t) + n_i(t)$$

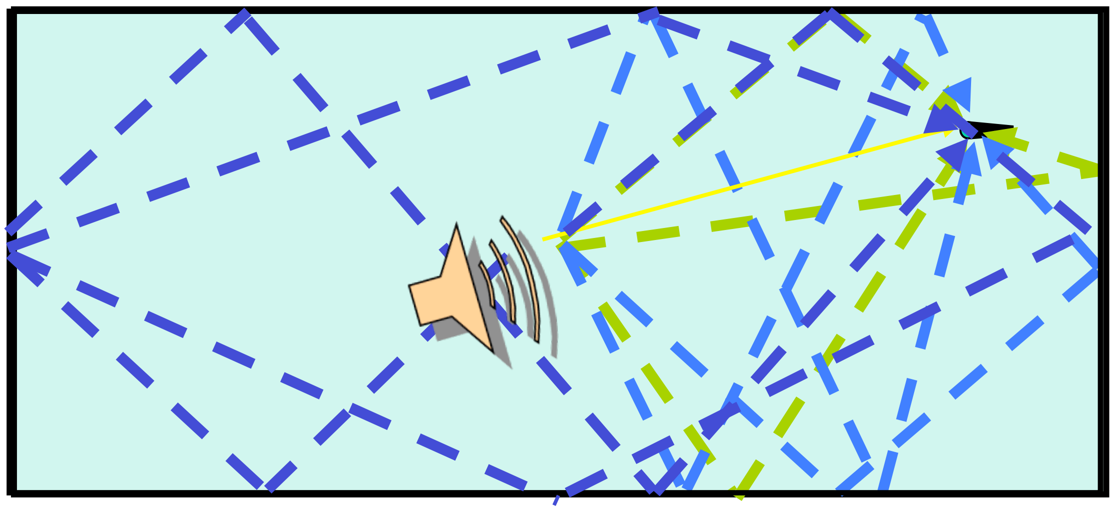
$h_i(\mathbf{s},t)$= acoustic **impulse response** for the given set of positions of the source and the microphone



Room acoustics: reverberation

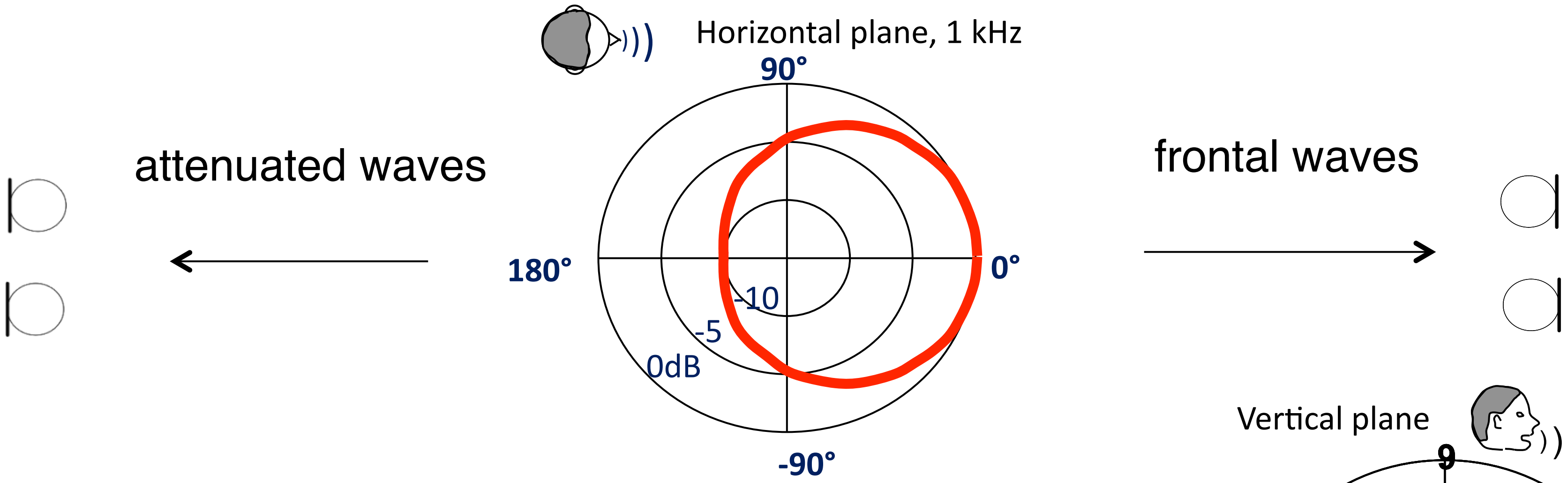
- The reverberation phenomenon is due to reflections from surfaces and diffusion and diffraction by objects inside the room
- It differs with the positions of source and listener (or microphone)
- For each source-microphone position, and source orientation (except in the omni-directional case), a different impulse response



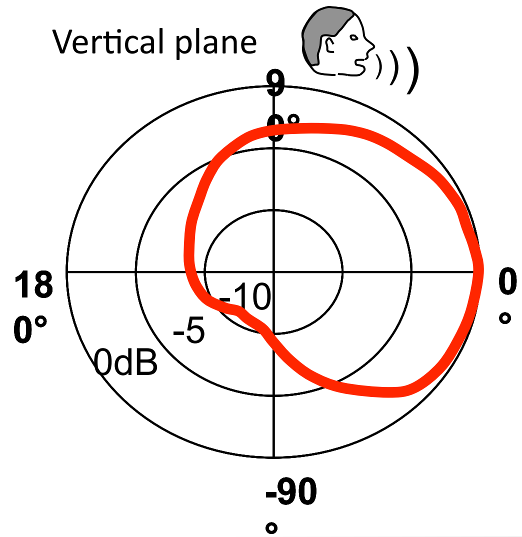


Source directivity

- Real sources (e.g. speakers) are not ideally omnidirectional
- Speakers have a distinct directivity pattern



- Frontal pairs: prevalence of direct energy
- Other pairs: prevalence of indirect energy



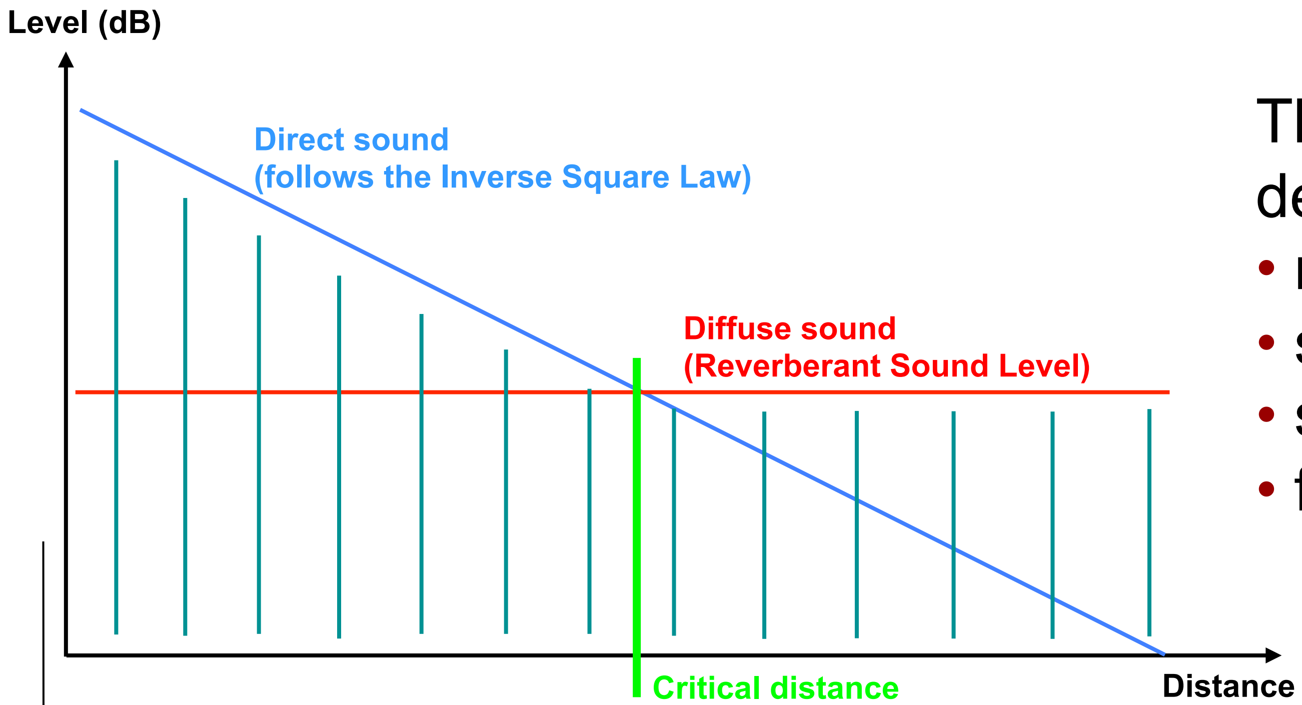
Maurizio Omologo – omologo@fbk.eu

Reverberation time and critical distance

Reverberation time T60: time required for a decay of 60 dB of intensity for a sound abruptly inerrupted.

Critical distance: distance from the sound source at which the direct and reflected sound intensities are equal.

Beyond the radius of critical distance, **Direct to Reverberant Ratio** (DRR) is negative (except for sound onset)



The critical distance depends on:

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

Maurizio Omologo – omologo@fbk.eu

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: $\boxed{r > \frac{2L^2}{\lambda}}$

where r is the distance to the array,

L is the length of the array, and

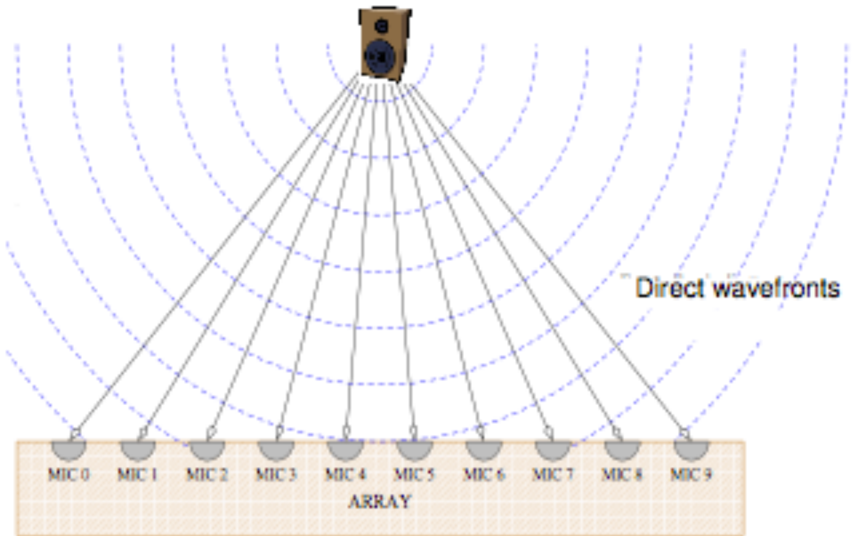
λ is the wavelength of the arriving wave

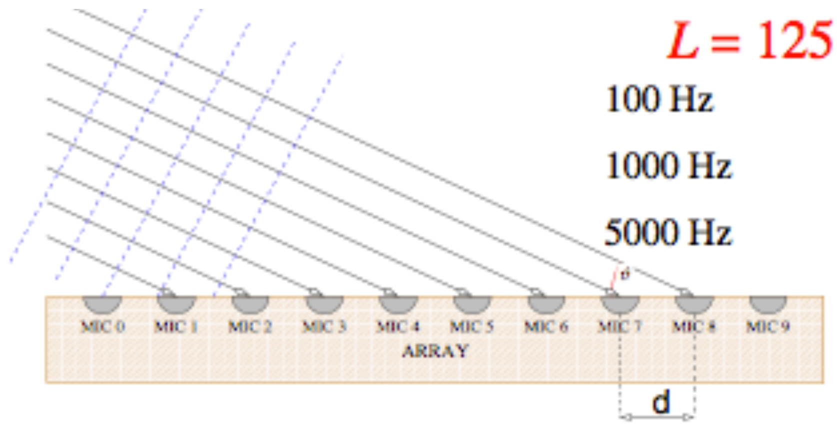
$L = 25\ cm$

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

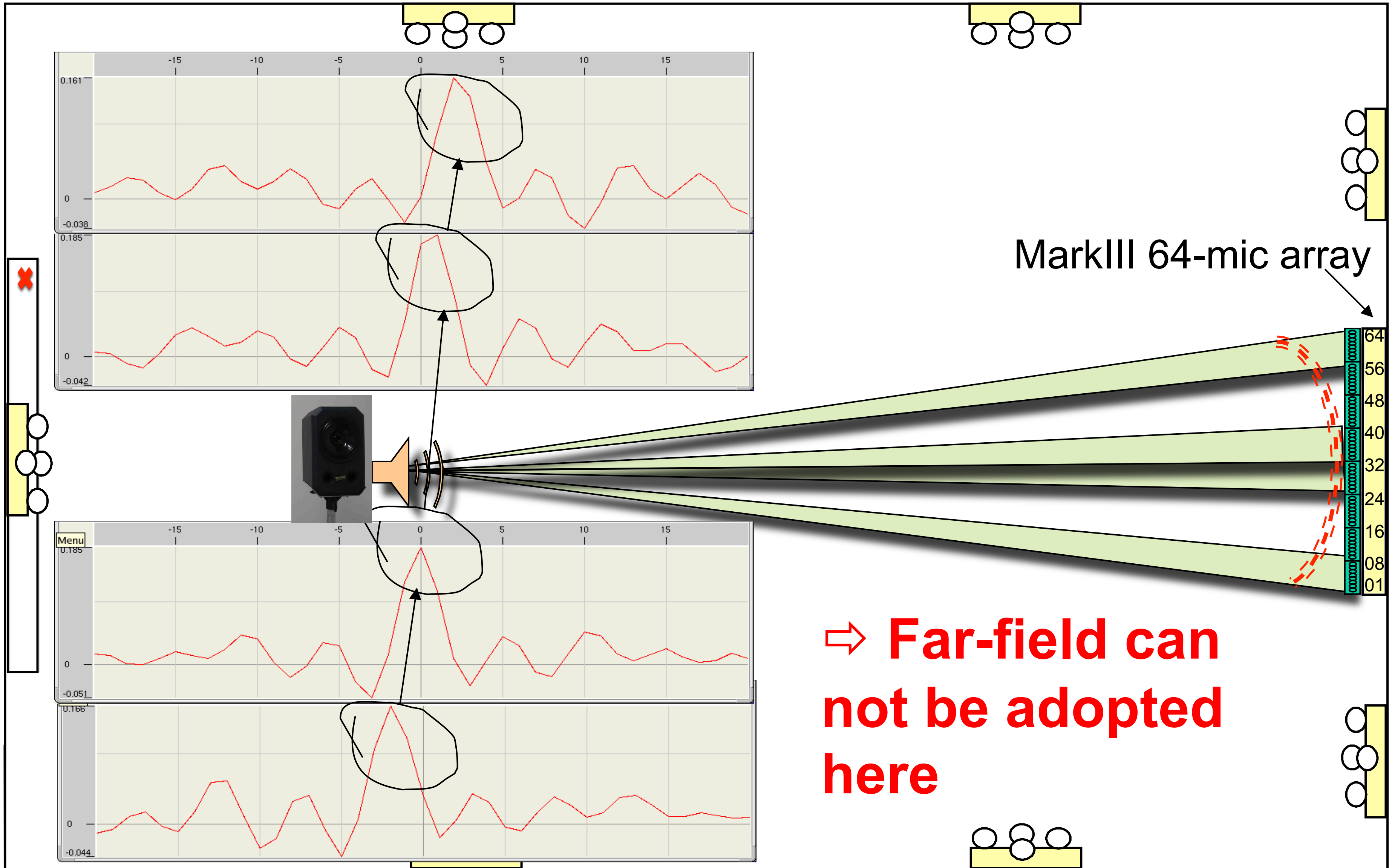
$L = 125\ cm$

$f =$	$r >\sim$
100 Hz	92 cm
1000 Hz	9.2 m
5000 Hz	46 m





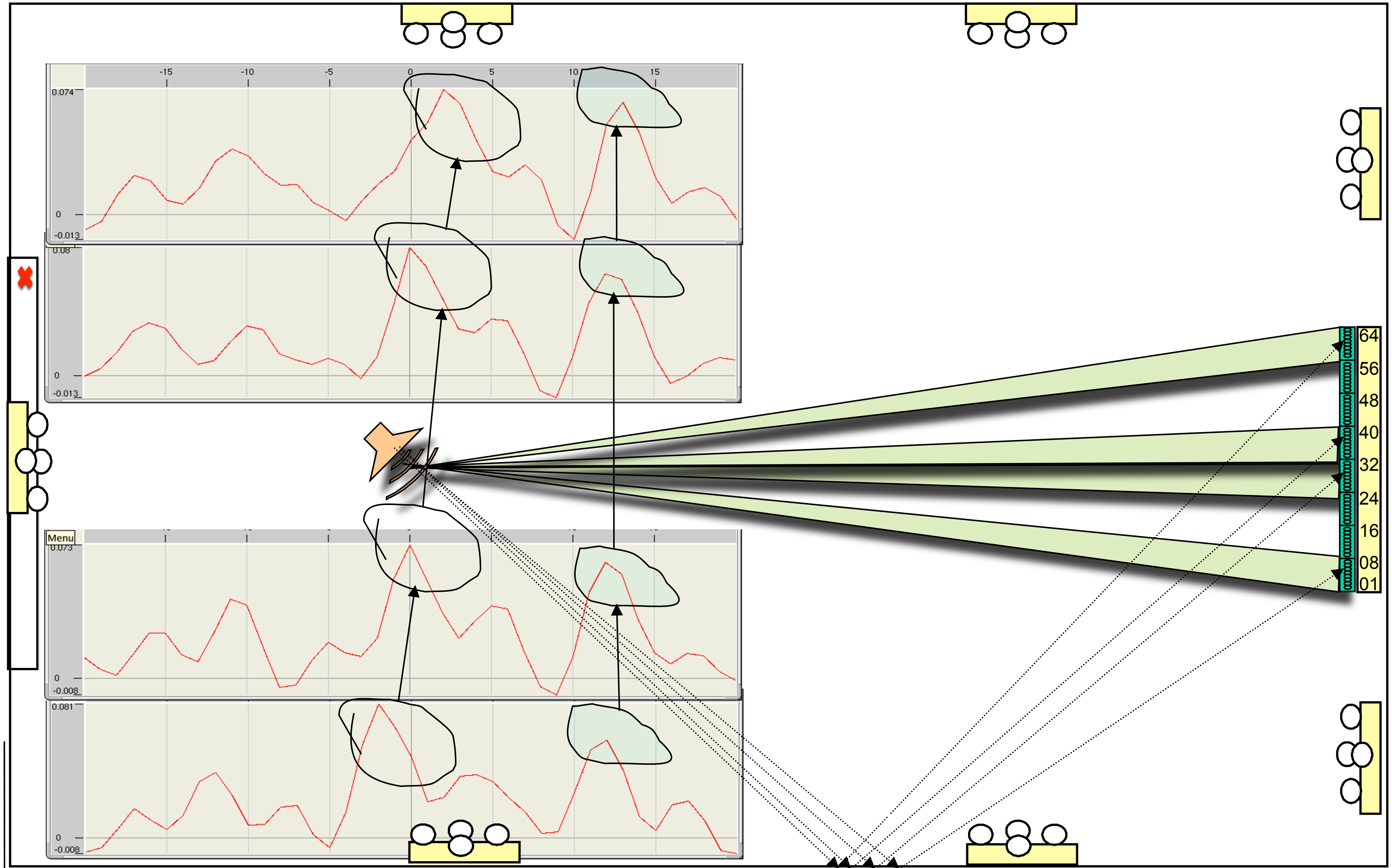
Loudspeaker oriented towards the array: GCC-PHAT analysis



⇨ **Far-field can not be adopted here**

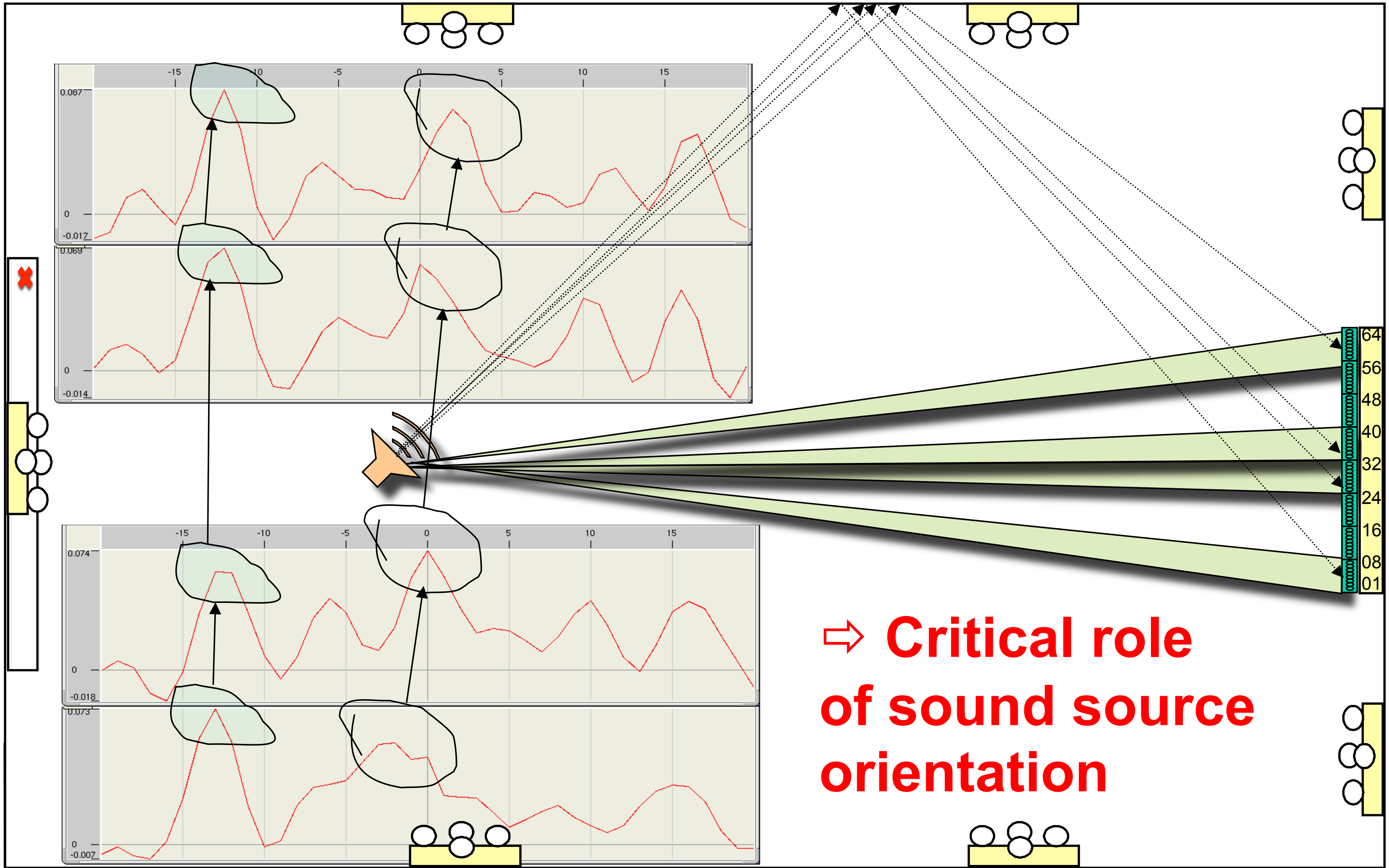
Maurizio Omologo – omologo@fbk.eu

Loudspeaker oriented to the right: GCC-PHAT analysis



Maurizio Omologo – omologo@fbk.eu

Loudspeaker oriented to the left: GCC-PHAT analysis



⇨ **Critical role of sound source orientation**

Maurizio Omologo – omologo@fbk.eu

The noise field at the microphones

The noise field results from contribution of different sources (unknown number, position and characteristics). We can distinguish among the following characteristics:

	vs	
Additive		Convolutional
Coherent field		Diffuse field
Point source		Spatially distributed source
Narrowband		Wideband
Uncorrelated		Correlated (with speech)
Stationary (in time and space)		Non-stationary
Known		Unknown

The complexity of acoustic source location and other multi-channel processing tasks can depend on the characteristics of the environmental noise and, in general, on SNR and DRR at each microphone.

Magnitude Square Coherence: experiments in the CHIL room at FBK

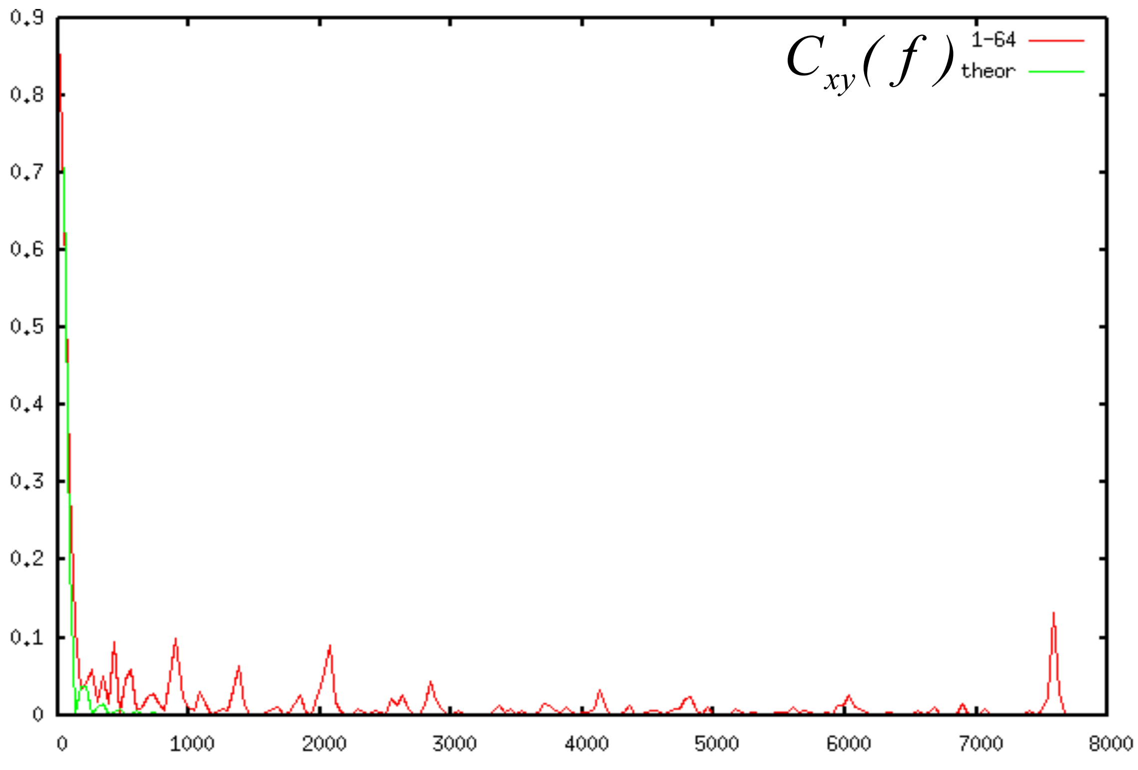
$$\gamma_{xy}(f) = \frac{P_{xy}(f)}{\sqrt{P_{xx}(f)P_{yy}(f)}}$$

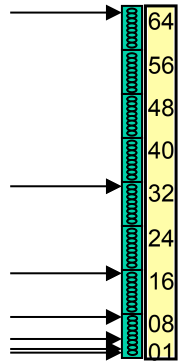
$$C_{xy}(f) = \frac{\left|P_{xy}(f)\right|^2}{P_{xx}(f)P_{yy}(f)}$$

For perfectly diffuse noise (spherically iso-tropic) and omnidirec-tional microphones:

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

[see Jacobsen, JASA 2000, and Brandstein-Ward, 2001]





Spatial coherence of background noise by means of different microphone pairs of the Mark III array. Comparison with the theoretical coherence for perfectly diffuse noise.

- Good feature for contexts characterized by stationary background noise and by no active acoustic sources diffusing spatially coherent fields
- … and also for calibration purposes

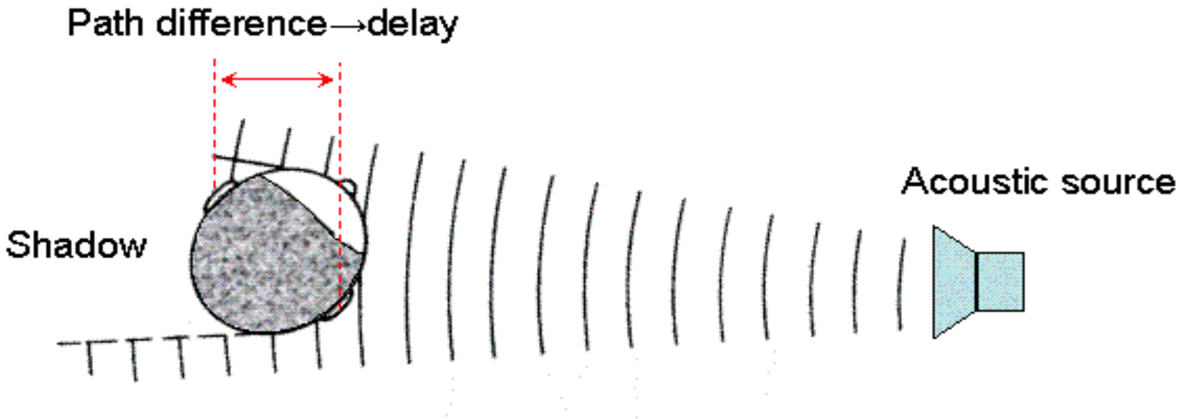
Single speaker localization and tracking

Binaural sound localization

The hearing system uses mainly two clues for estimating the ***Direction Of Arrival*** (DOA) of the sound generated by an acoustic source:

- Interaural Intensity Difference (IID)
- Interaural Time Difference (ITD)

- ITD is considered the most important hint for sound localization.
- IID is mainly useful above 1500 Hz, where the acoustic shadow produced by the head becomes effective.



see [Blauert, 1983/1997] for a good overview

Human perception ⟺ ITD
Automatic source location ⟺ ***Time Difference of Arrival*** (TDOA)

Acoustic source location: common approaches and techniques

1. Indirect Methods ➡ <u>dual-step procedures</u>
2. Direct Methods ➡ <u>single-step procedures</u>

a. TDOA estimation
b. Apply geometry

1. Memoryless solutions ➡ frame-by-frame independent estimation of the source position
2. Memory-based solutions ➡ <u>regularize a sequence of positions</u> on a temporal interval longer than one frame

Kalman, Particle filtering, Gauss, etc

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

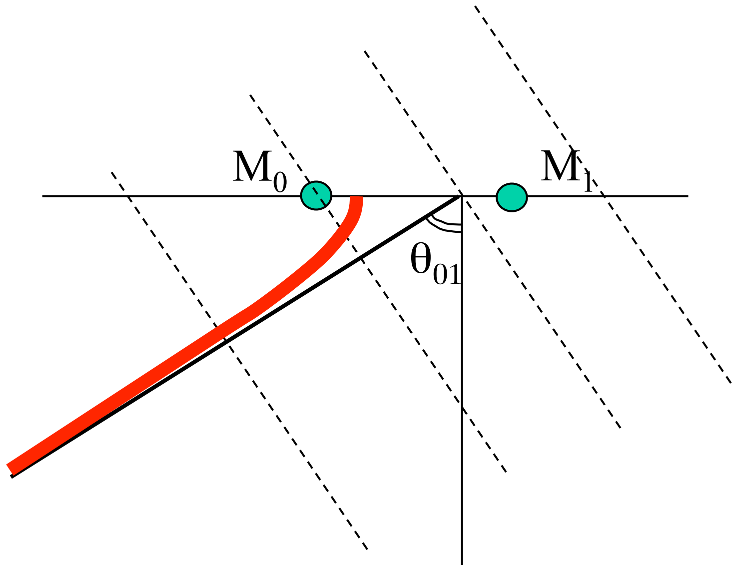
Eigenanalysis, MUSIC, ESPRIT, BSS, etc

High-resolution spectral estimation based on the signal correlation matrix and other techniques

Speaker Location (SLOC) and Tracking

- Most of the research activities since 1990
- Early technologies inspired by binaural sound source localization (mostly based on interaural time difference)
- The most critical issue: derive a Time Difference of Arrival with *high accuracy* from a microphone pair input

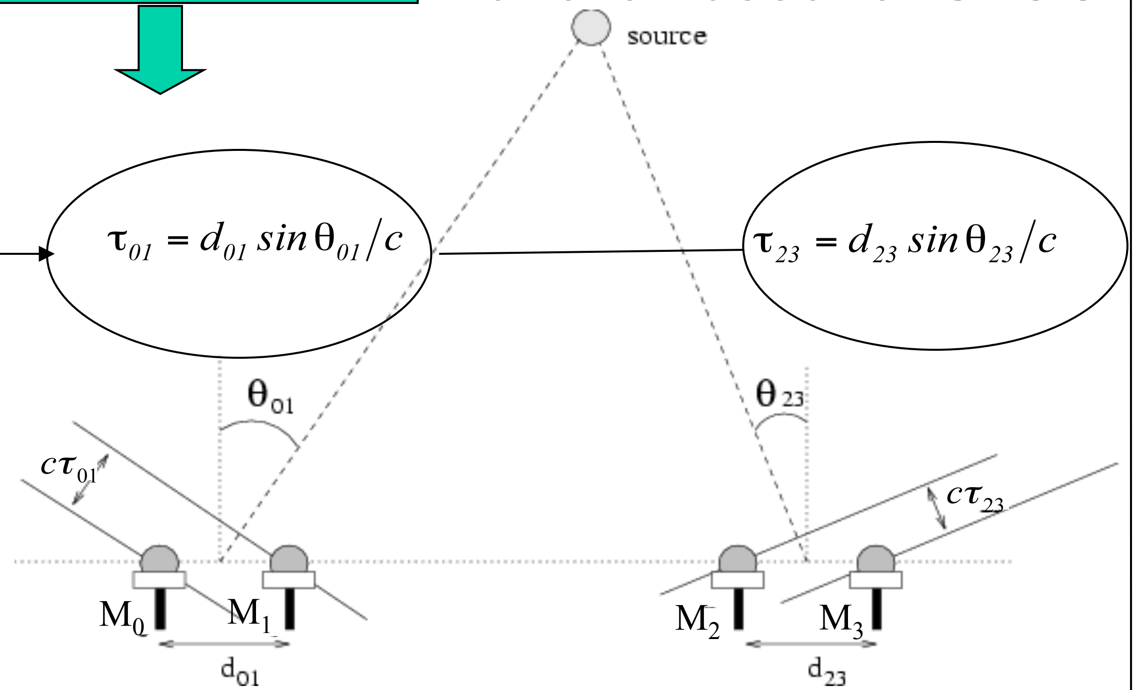
Trivial two-step solution based on two microphone pairs



$$\tau_{01} = \arg\max_{|\tau|<d_{01}/c}\left[gcc_{01}(\tau)\right]$$

$$\theta_{01} = \arcsin\left(\frac{c\,\tau_{01}}{d_{01}}\right)$$

GCC-PHAT is the most common correlation function used for SLOC



a. Estimate the two delays
b. Cross the resulting directions

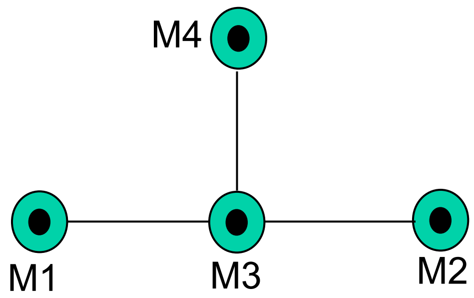
TDOA error statistics vs location accuracy depends on geometry

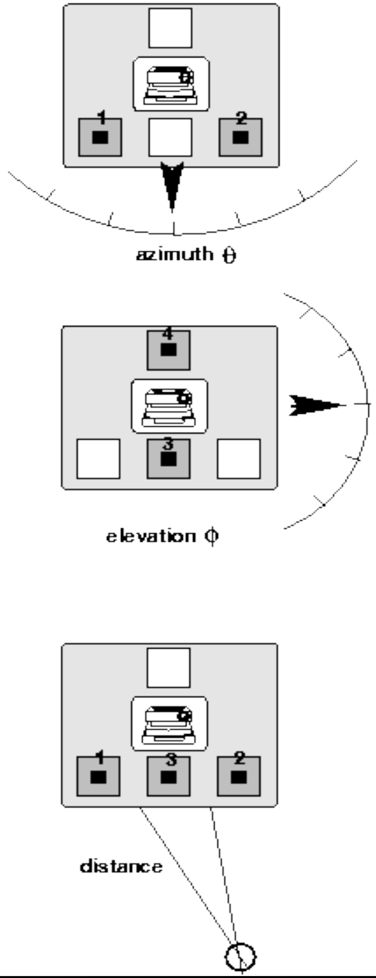
Speaker Location (SLOC) and Tracking

- ✓ *Most of the research activities since 1990*
- ✓ *Early technologies inspired by binaural sound source localization (mostly based on interaural time difference)*
- ✓ *<u>The most critical issue:</u> derive a Time Difference of Arrival with high accuracy from a microphone pair input*
- o Other major issues:
  - Microphone array **Geometry**
  - **Quantity** and **Quality** of the microphones
  - Characteristics of **Environmental Noise** and **Reverberation**
  - Number of **Active Sources** and related spectral contents
  - **Head Orientation** (or radiation pattern of a generic source)
  - Combine Speaker Location, with **Speaker ID**, and **Acoustic Event Detection**
  - System **Promptness** (even with short events, overlapping each other)

Use of a reverse T-shaped array geometry

Bearing direction and range estimation can be obtained with a reversed T-shaped microphone configuration





$$\theta = \arcsin\left(\frac{c \cdot delay_{12}}{d_{12}}\right)$$

$$\phi = \arcsin\left(\frac{c \cdot delay_{34}}{d_{34}}\right)$$

$$\left.\begin{matrix} delay_{13} \\ delay_{32} \end{matrix}\right\} \rightarrow \text{distance}$$

- o use of M1, M2, M3, for a 2D location (source assumed on the same plane):
  - ➢ M1-M2 for estimation of azimuth
  - ➢ M1-M3 and M3-M2 for distance
- o use of M3-M4 for elevation

Prototypes and products based on a reverse T-shaped array geometry

Prototype of speaker location and tracking realized at ITC-irst in 1994

Device for videoconferencing produ-ced by AETHRA (Italy) since 1999

- Developed under EC DIMUS project (surveillance of metro stations)
- Since 1997, automatic source location embedded in products for videoconferencing (e.g., PictureTel, Polycom)
- Reverse T-shaped geometry was the most commonly used

Global Coherence 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:

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

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

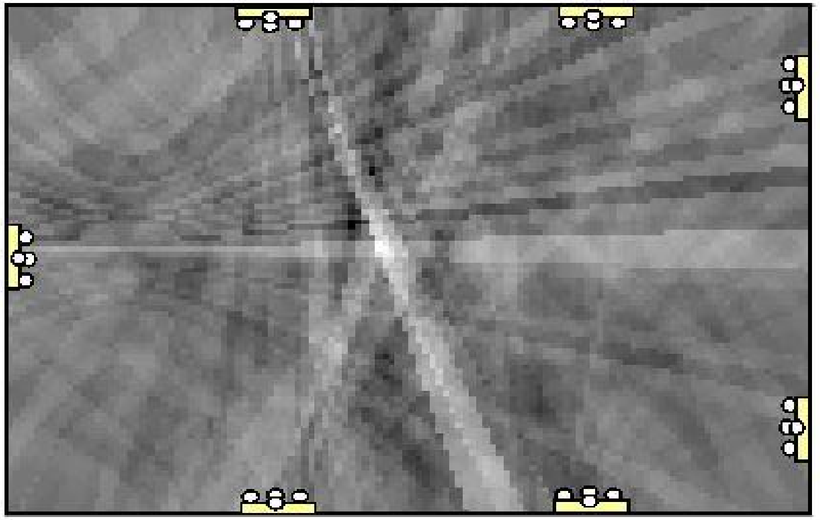
$$\Rightarrow \hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s}} GCF(t,\mathbf{s})$$

Pros: GCF provides a sharper peak than alternative approaches, with a consequent decreased sensitivity to noise and reverberation. Moreover, it is a direct single-step method.

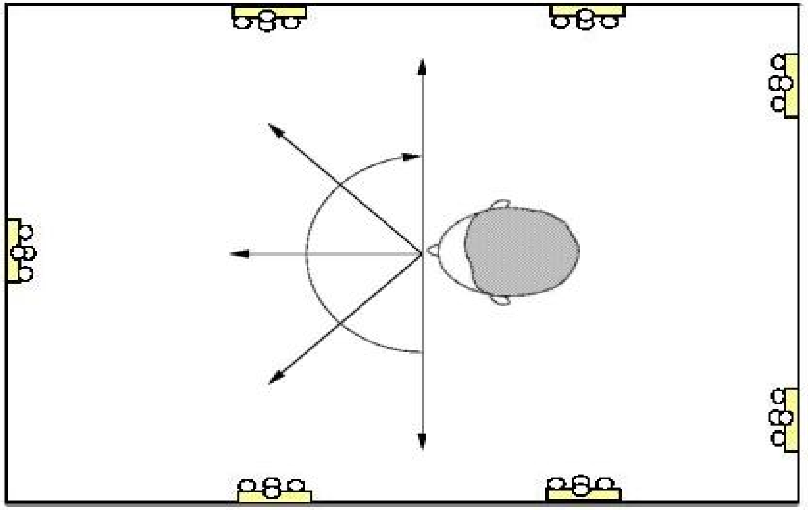
Cons: Possible weakness depending on geometry, room acoustics, speaker position, head orientation, etc.

*** References in the literature [see Brandstein-Ward 2001] often use the term SRP-PHAT to indicate the above described GCF technique.**

Use of GCF to estimate Head Orientation



Example of 2D GCF in a real room

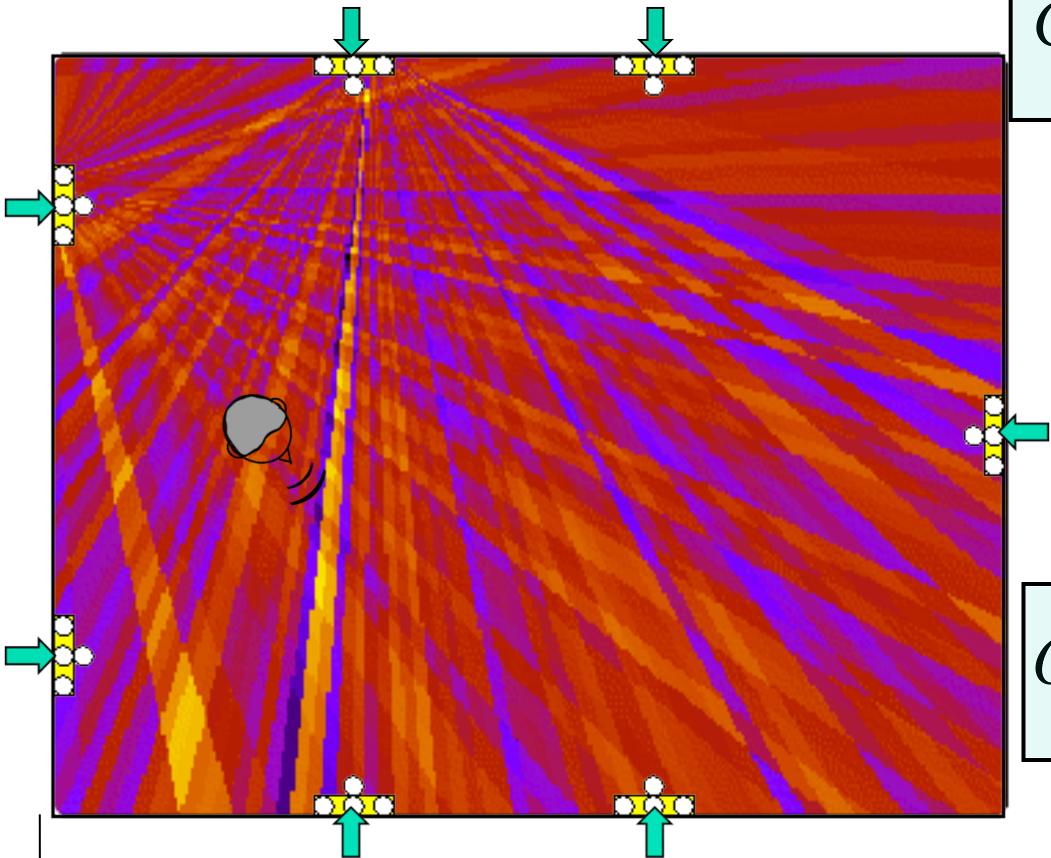


According to head orientation the contribution of the various micro-phone pairs have different strength

- 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)

From GCC-PHAT to Global Coherence Field (GCF) and Oriented GCF

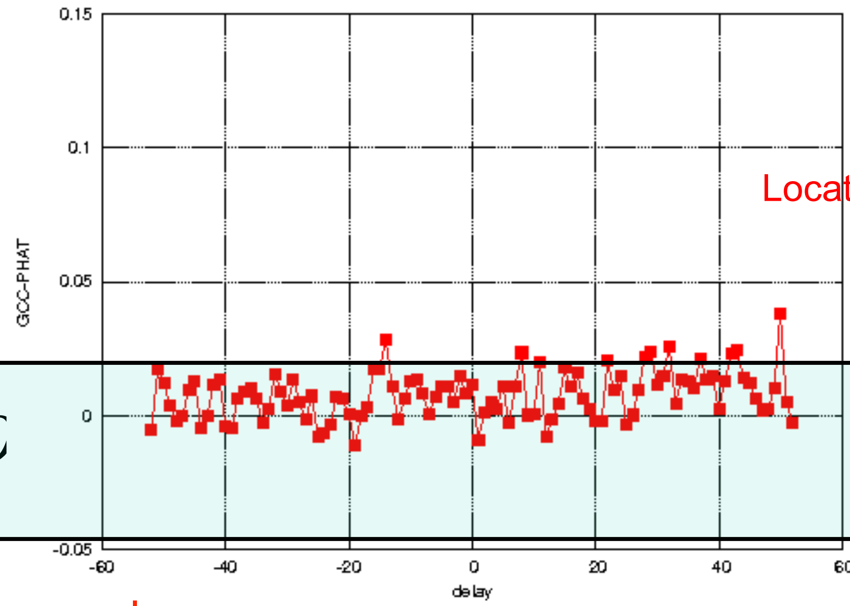
$$gcc_{01}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{Y_0(\omega)Y_1^*(\omega)}{|Y_0(\omega)Y_1^*(\omega)|}e^{j\omega\tau}d\omega$$



Given M_P microphone pairs:

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

$\delta_{ik}(\mathbf{s})$ = theoretical delay



)

Location

C ... lj)

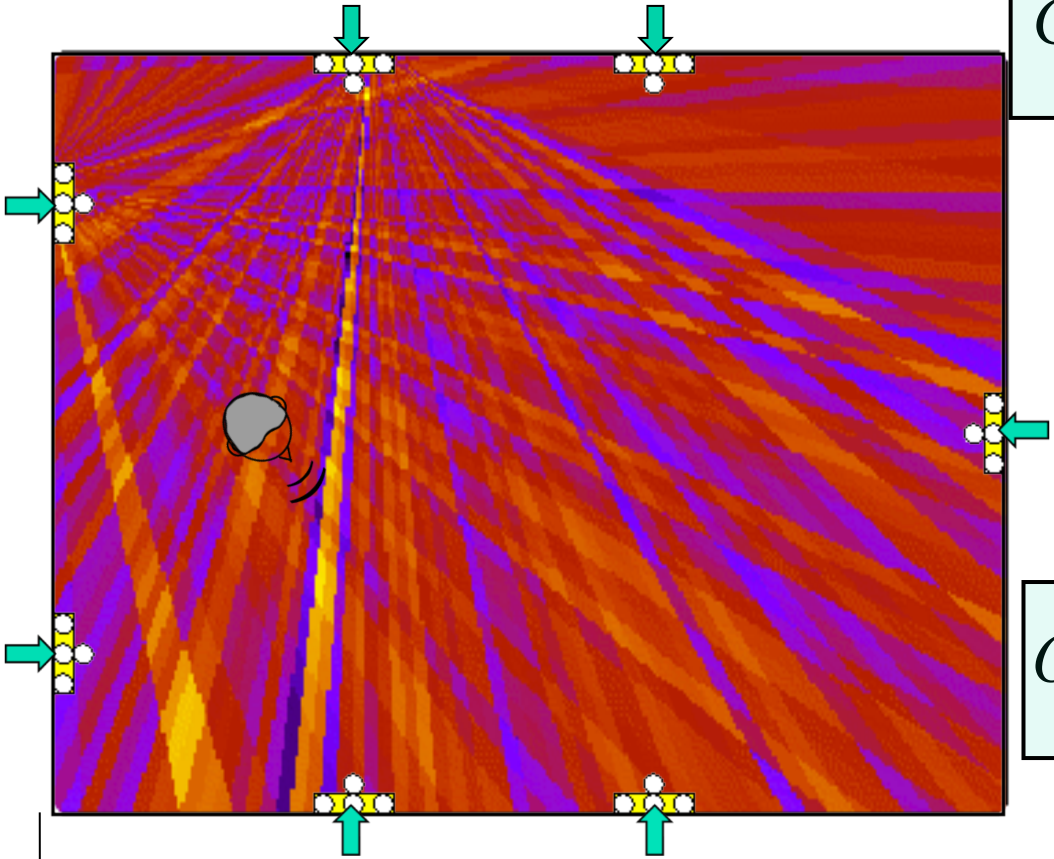
$$\hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s},j} OGCF_j(t,\mathbf{s})$$

Location+Orientation

More details on GCF in [De Mori 1998] and, in terms of SRP-PHAT, in [Di Biase et al. 2001]. As for OGCF, see [Brutti et al. 2005]. Extended to multiple speaker location, see [Brutti et al. 2008].

From GCC-PHAT to Global Coherence Field (GCF) and Oriented GCF

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



Given M_P microphone pairs:

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

$\delta_{ik}(\mathbf{s})$ = theoretical delay

$$\hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s}} GCF(t,\mathbf{s})$$

Location

for every direction j:

$$OGCF_j(t,\mathbf{s}) = \sum_{l=0}^{L-1} GCF_{\Omega_l}(t,Q_l)w(\theta_{lj})$$

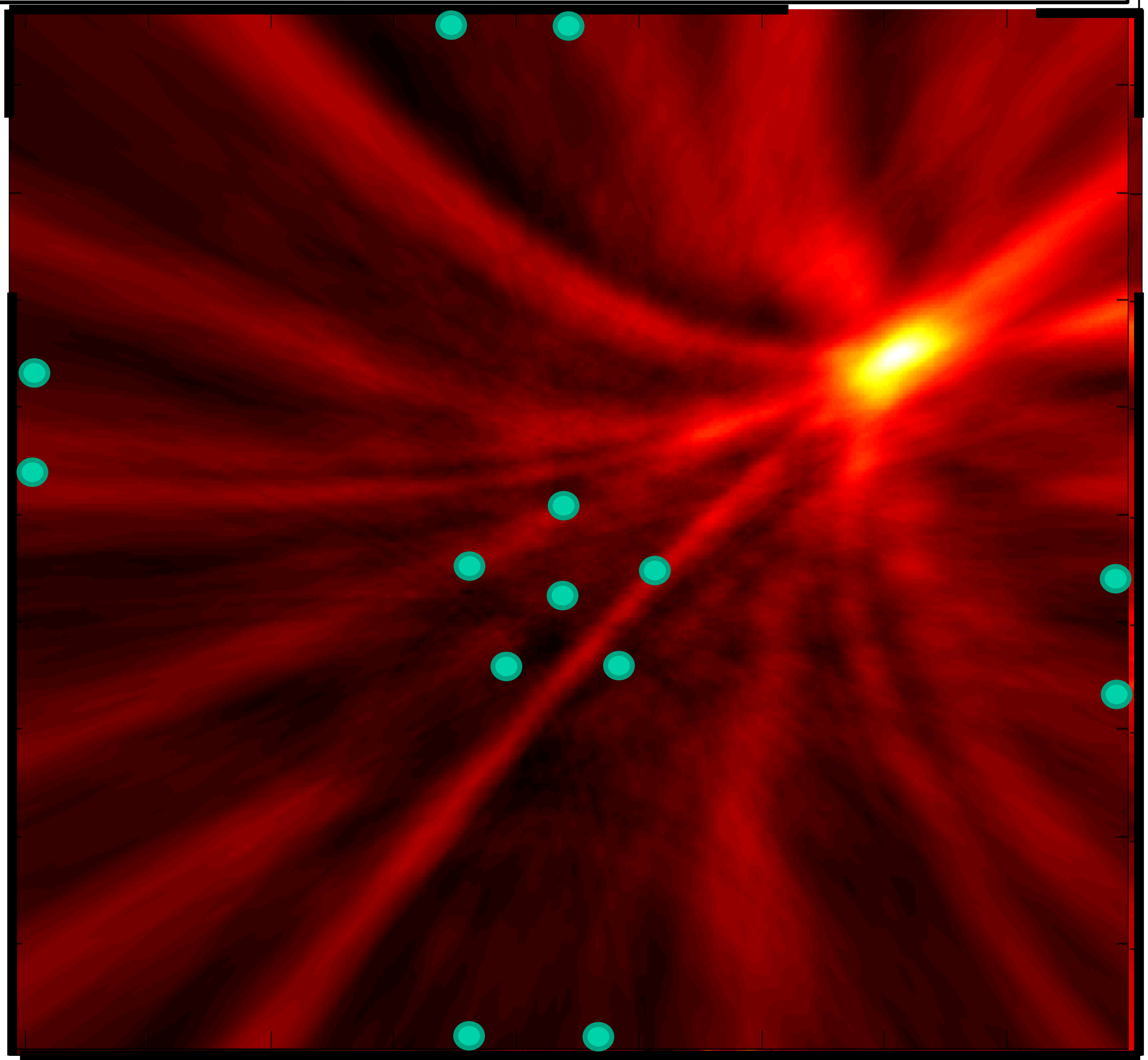
$$\hat{\mathbf{s}}(t) = \arg\max_{\mathbf{s},j} OGCF_j(t,\mathbf{s})$$

Location+Orientation

More details on GCF in [De Mori 1998] and, in terms of SRP-PHAT, in [Di Biase et al. 2001]. As for OGCF, see [Brutti et al. 2005]. Extended to multiple speaker location, see [Brutti et al. 2008].

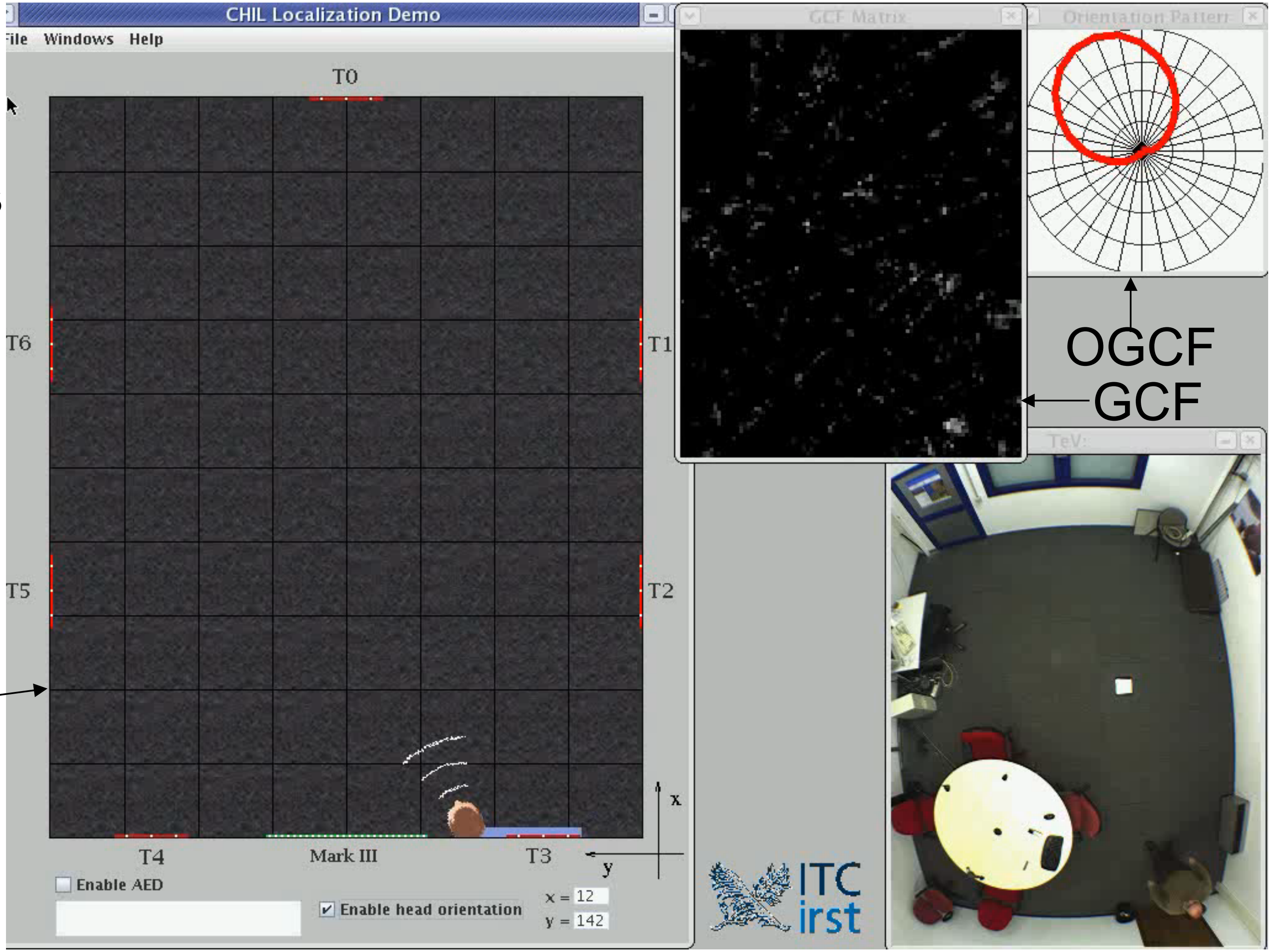
DIRHA - Acoustic Maps: ceiling-array vs wall-mic pairs

- Acoustic scene observed in the living room of the ITEA apartment
- Comparison between GCFs:
  - based on a six-microphone array installed on the ceiling
  - based on four microphone pairs on the walls
- Combination of GCFs gives more precision both in 3D location and head orientation



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

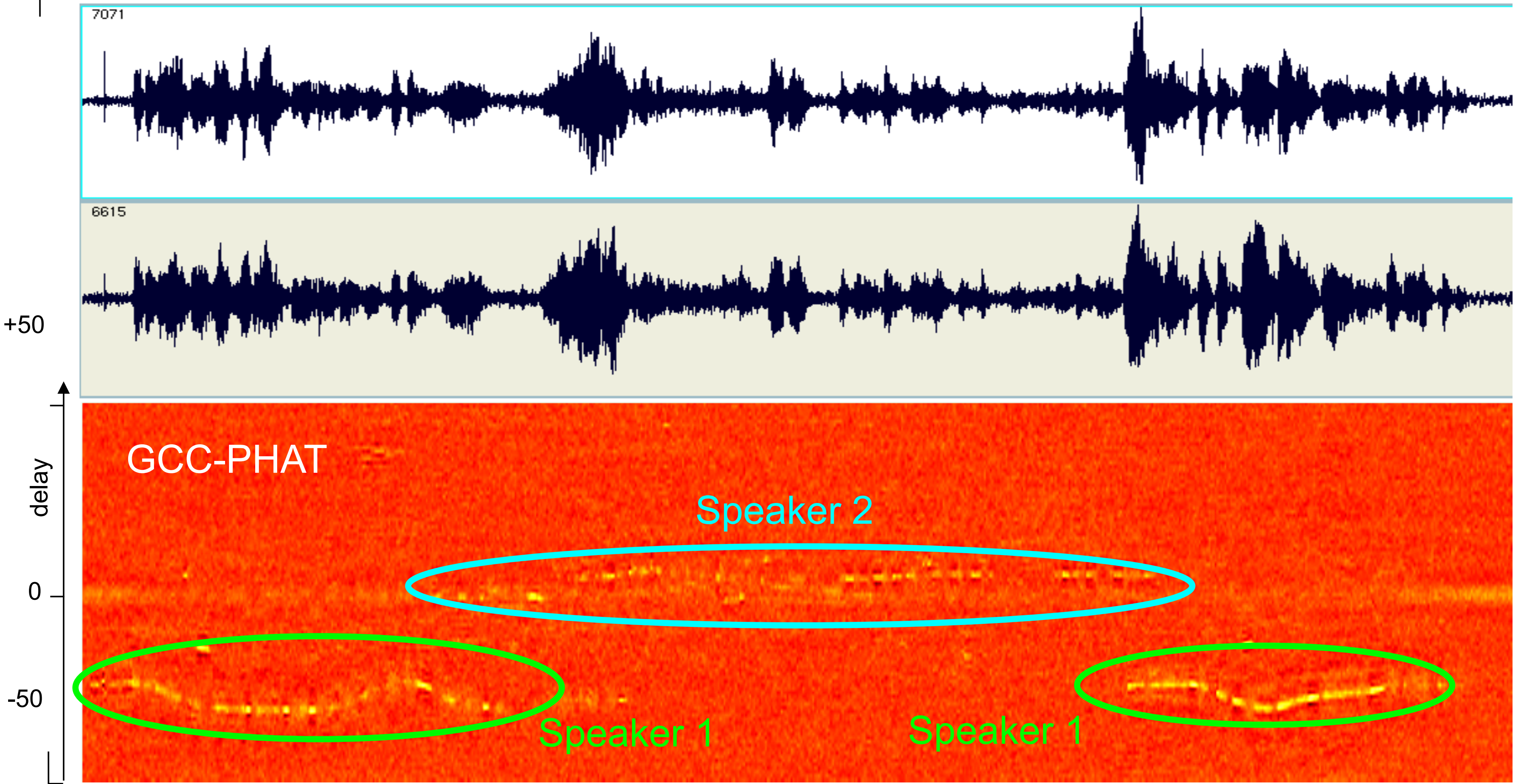


Map of the CHIL room at FBK-irst

Multiple speaker localization and tracking

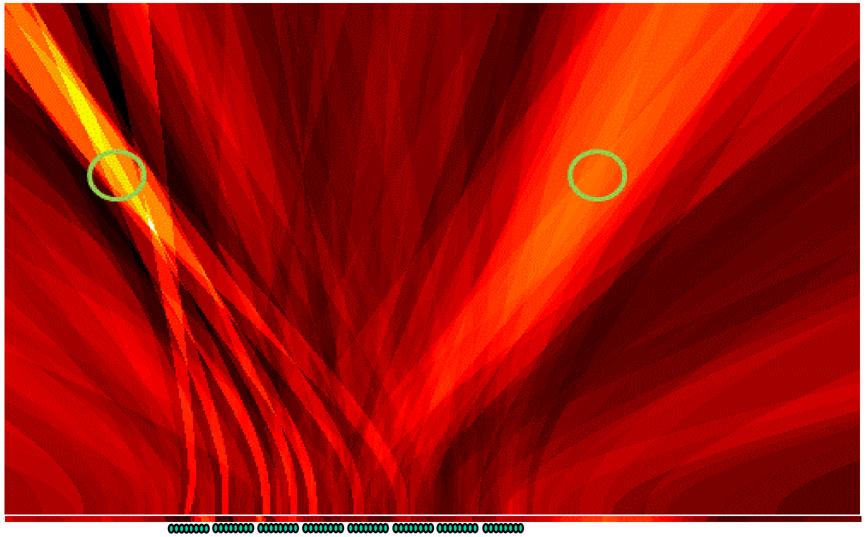
Maurizio Omologo – omologo@fbk.eu

Application of GCC-PHAT with two speakers



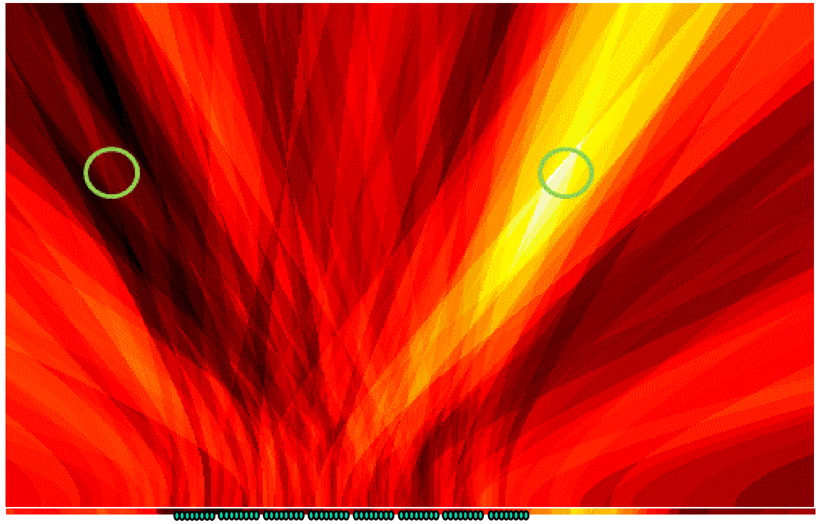
Maurizio Omologo – omologo@fbk.eu

Multi-step GCF de-emphasis to locate multiple speakers



Example of GCF in a real room given two individuals who are speaking simultaneously.

Different space positions are characterized by high GCF values. However, one can find the dominant speaker close to the upper left corner.



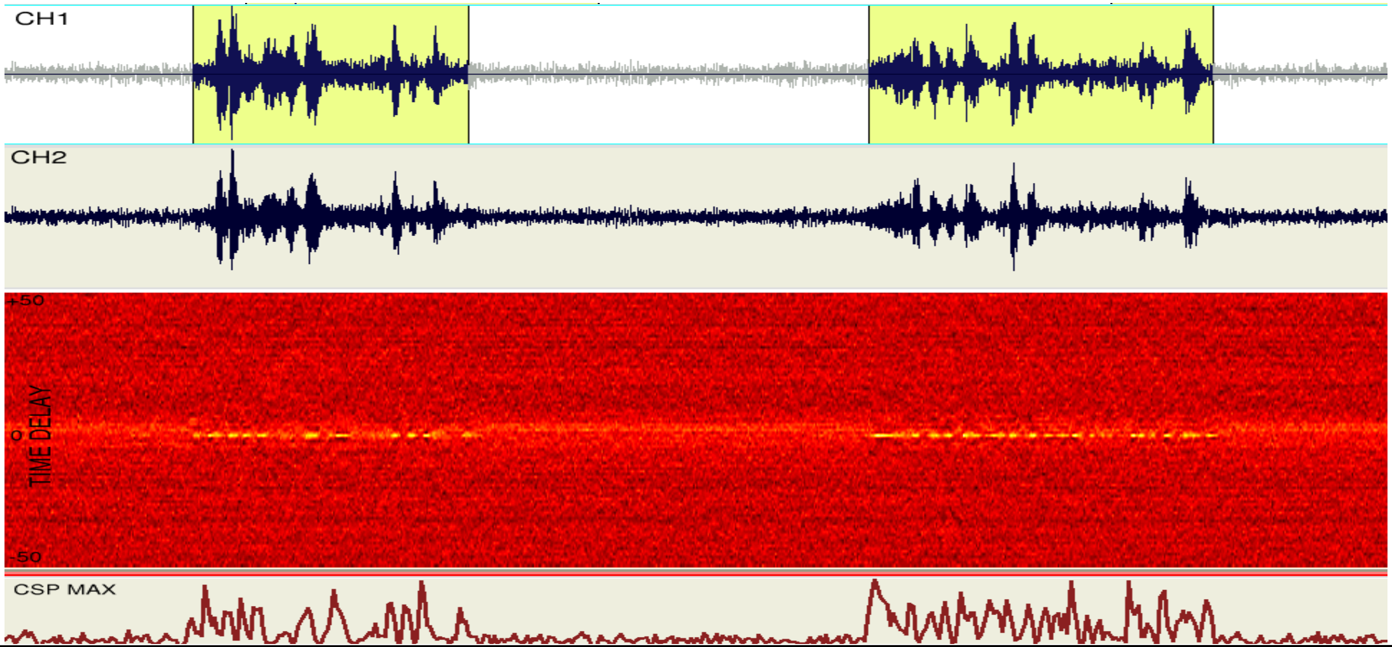
The new normalized GCF is used to look for a possible second active speaker.

GCF is then processed in order to remove contributions refered to the located speaker. GCF is also normalized (in order to apply again the same thresholding).

**more details can be found in Brutti et al., ICASSP 2008*

Speech Activity Detection (SAD) for Speaker Location and Tracking

- 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 CSP, or of GCF and 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:

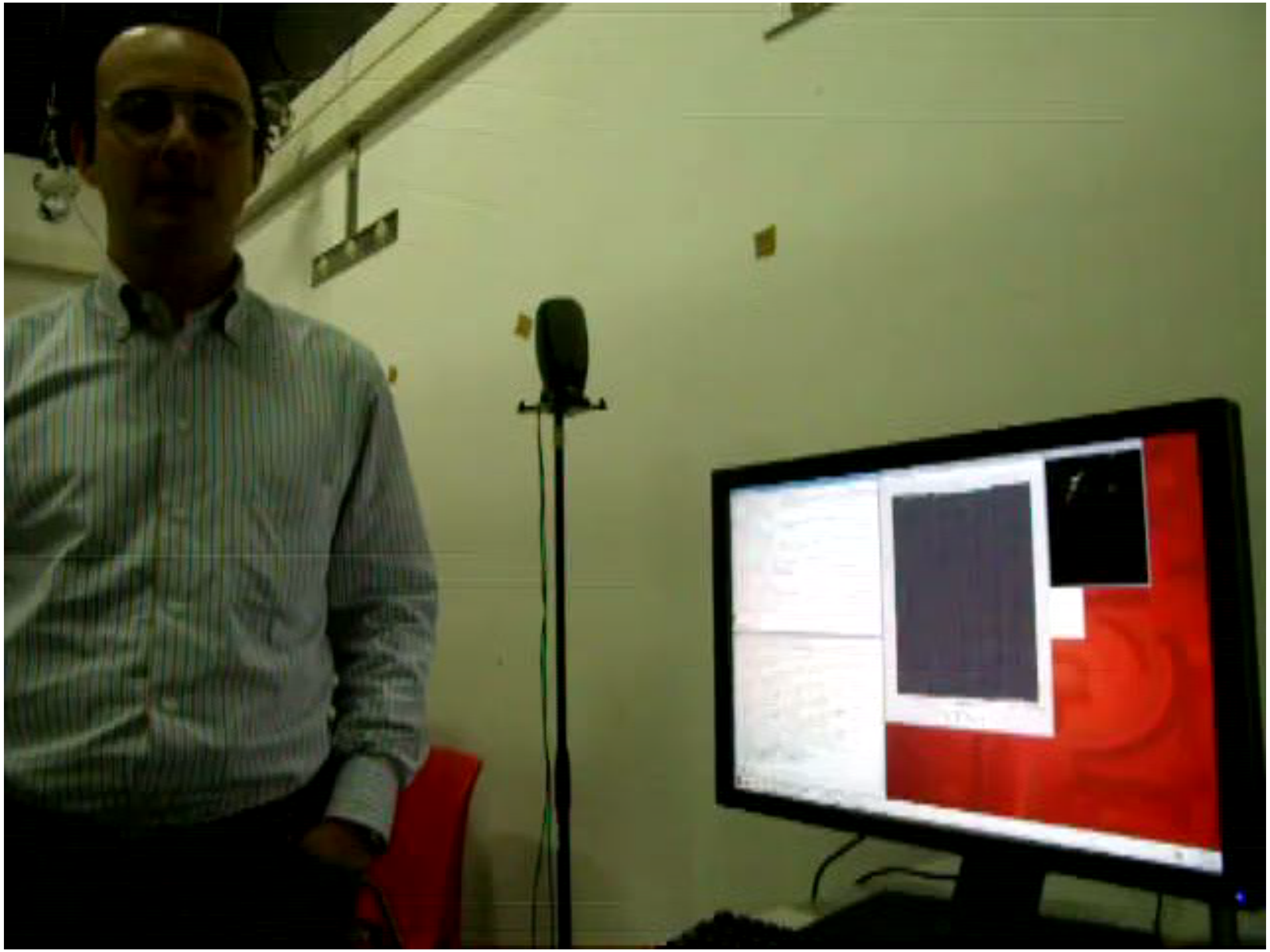


Multiple Speaker Tracking – 2 speakers

- 2-D real-time multi-speaker tracking based on 7 microphone triplets
- Use of particle filtering
- Filtering based on an embedded speech activity detector
- Use of GCF de-emphasis to filter different speakers

Multiple Speaker Tracking – 3 speakers

- 2-D real-time multi-speaker tracking based on 7 microphone triplets
- Use of particle filtering
- Filtering based on an embedded speech activity detector
- Use of GCF de-emphasis to filter different speakers



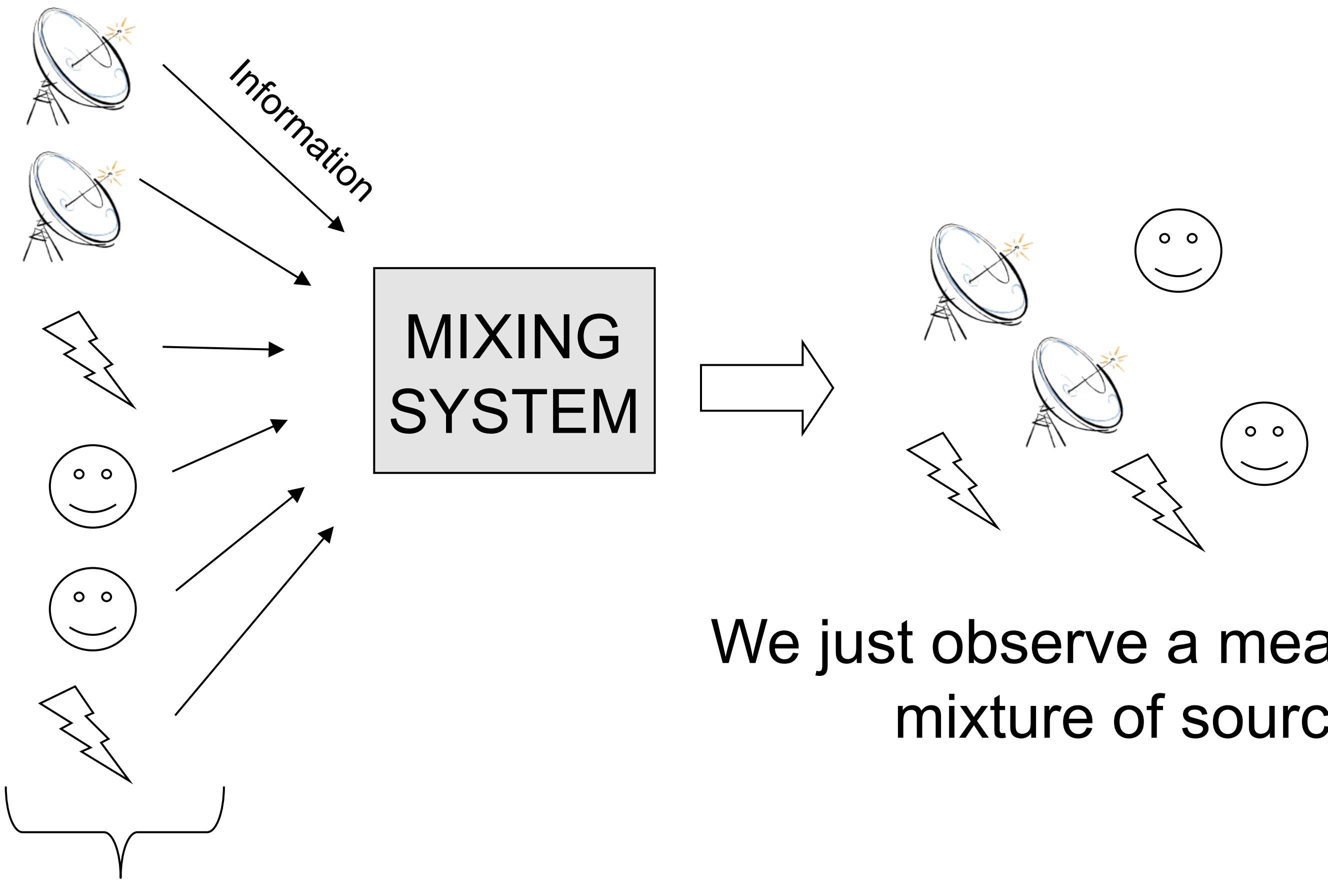
Source enhancement
Blind source separation and extraction

Maurizio Omologo – omologo@fbk.eu

Multi-channel source enhancement

- Target: improve the quality of the desired speech, by suppressing/ mitigating disturbances/distortions due to multiple sources in a real environment
- Multi-channel input ⟹ spatial filtering, source separation, interference cancellation
- Various enhancement approaches
  - Adaptive beamforming + postfiltering
  - Blind source separation (based on Independent Component Analysis)
  - Source-model techniques extended to distant-speech
- Source separation and extraction
  - Attractive solution, generally providing cues about source location
  - Non trivial and still open at theoretical level
  - With proper constraints ⟹ satisfactory performance

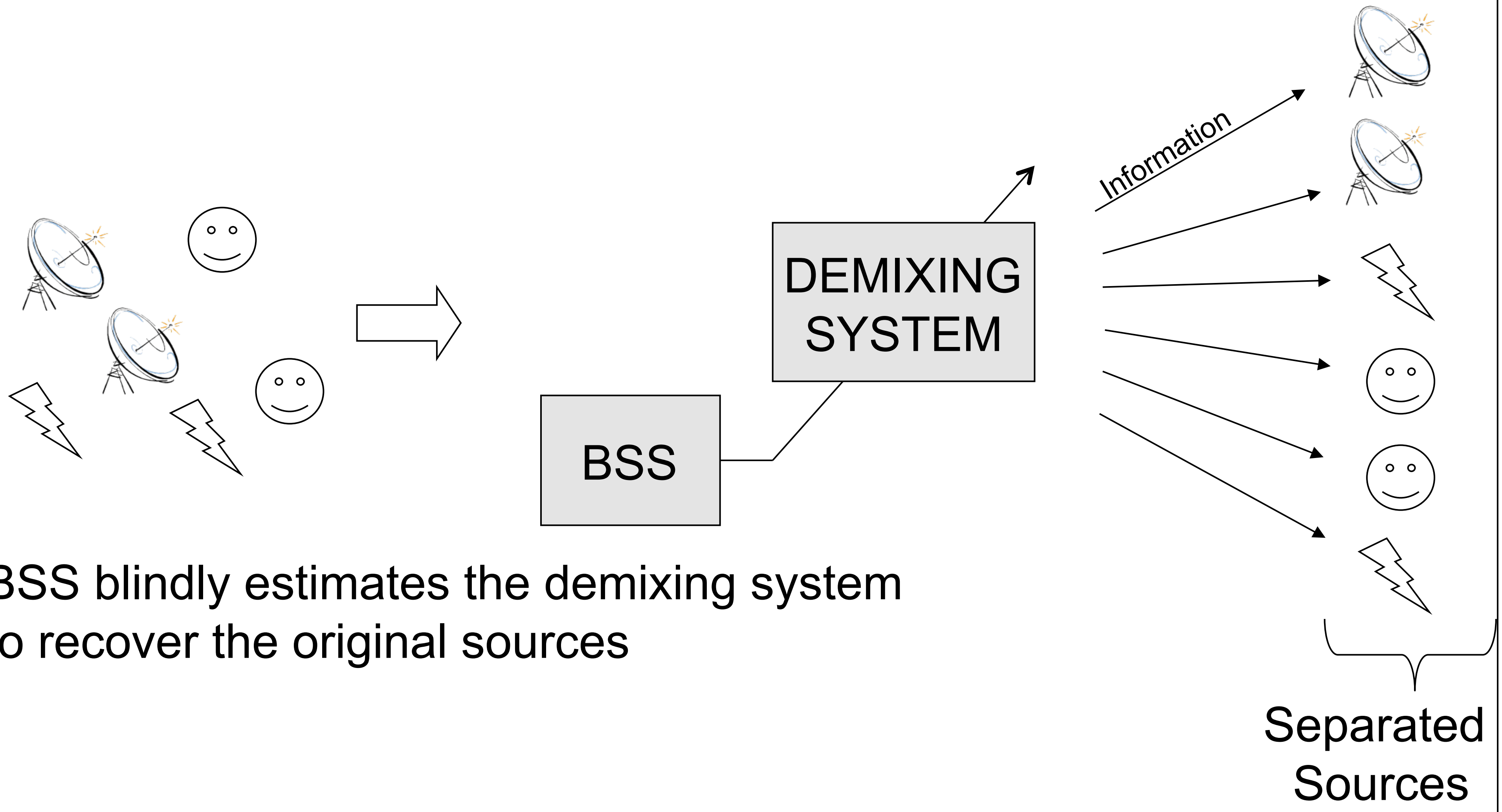
What is Blind Source Separation (BSS)?



We just observe a meaningless mixture of sources

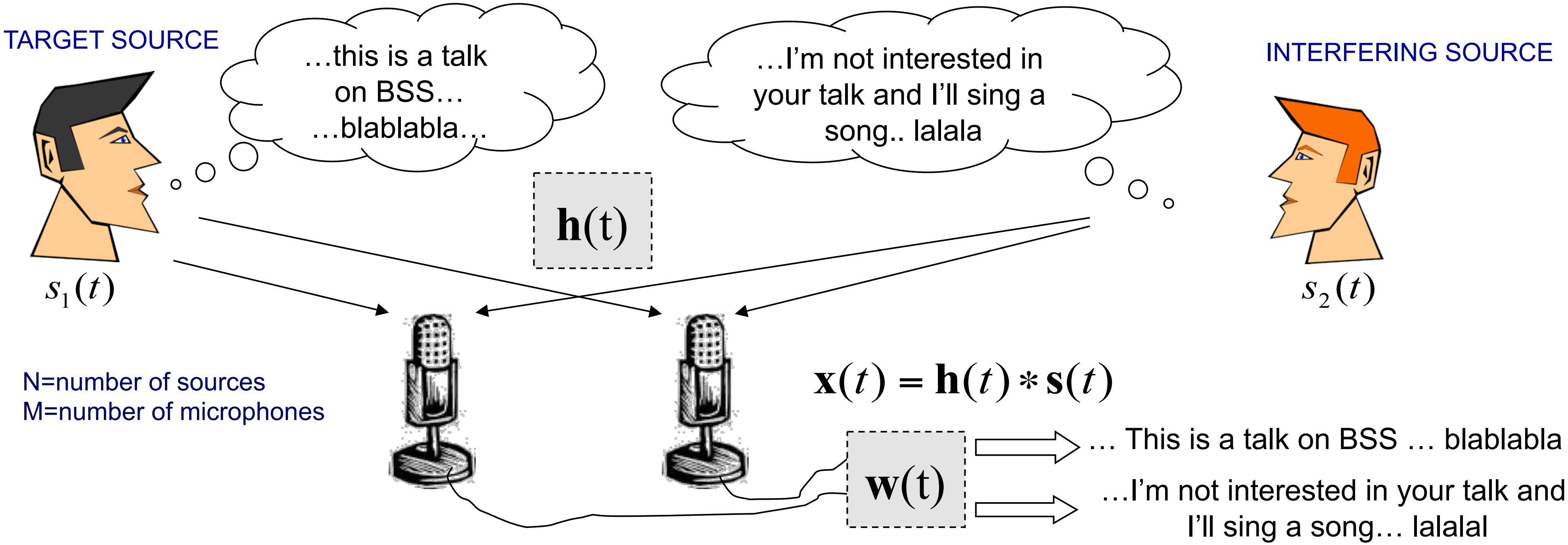
Independent Sources

Introduction to BSS (2/2)



BSS blindly estimates the demixing system to recover the original sources

Source separation for speech enhancement



N=number of sources
M=number of microphones

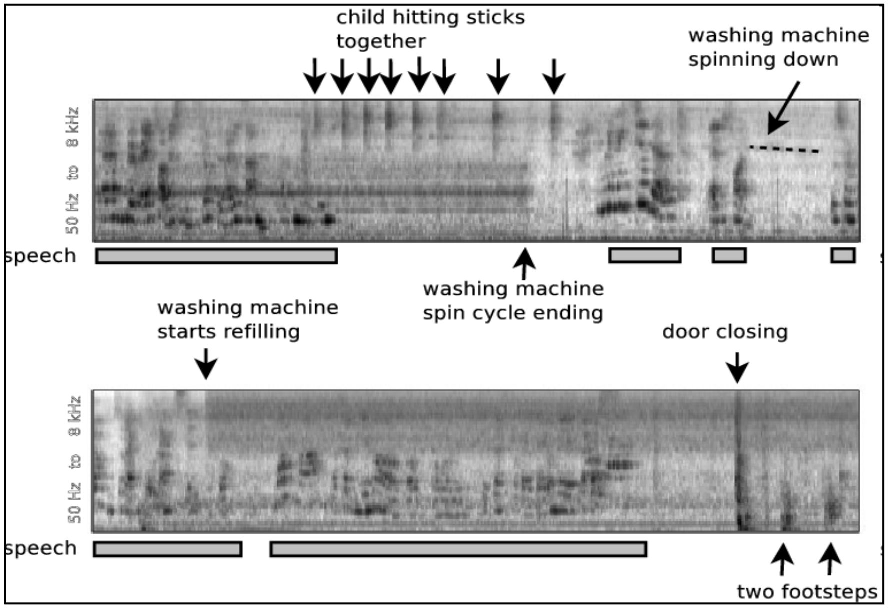
$$\mathbf{x}(t) = \mathbf{h}(t) * \mathbf{s}(t)$$

If N=M→ determined scenario

$$\mathbf{y}(t) = \mathbf{w}(t) * \mathbf{x}(t) \approx \mathbf{s}(t)$$

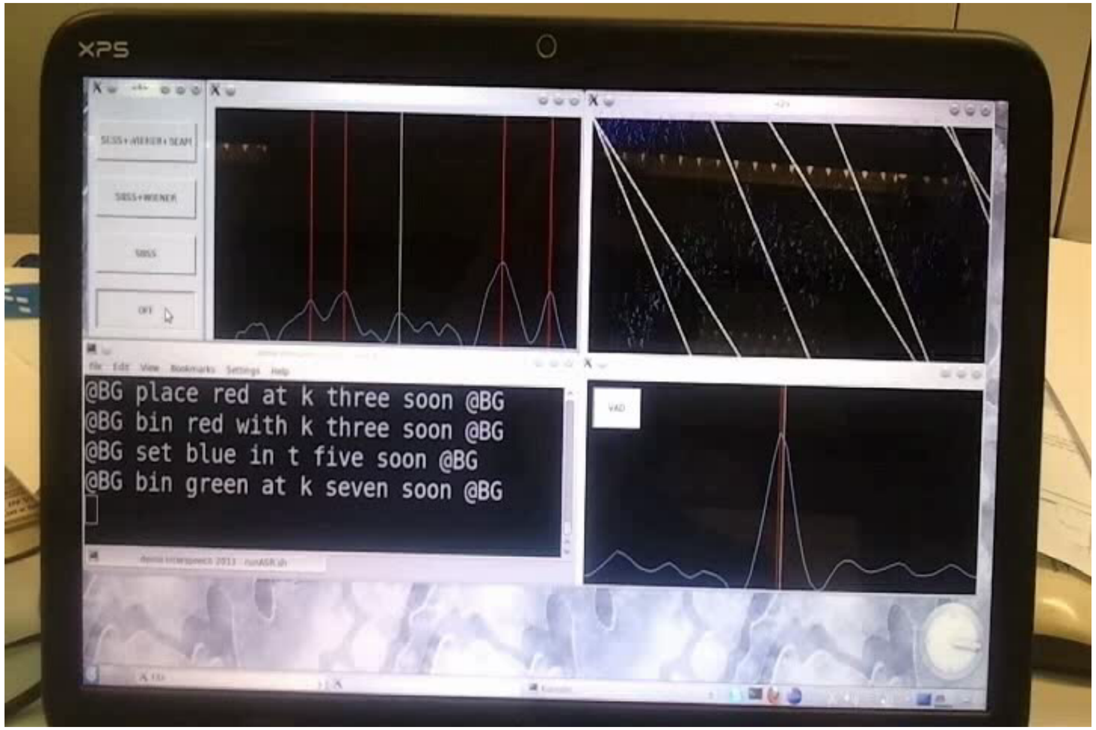
Maurizio Omologo – omologo@fbk.eu

CHIME challenge 2011



Real-time Blind Source Extraction + Speech Recognition

Demo presented at Interspeech 2011

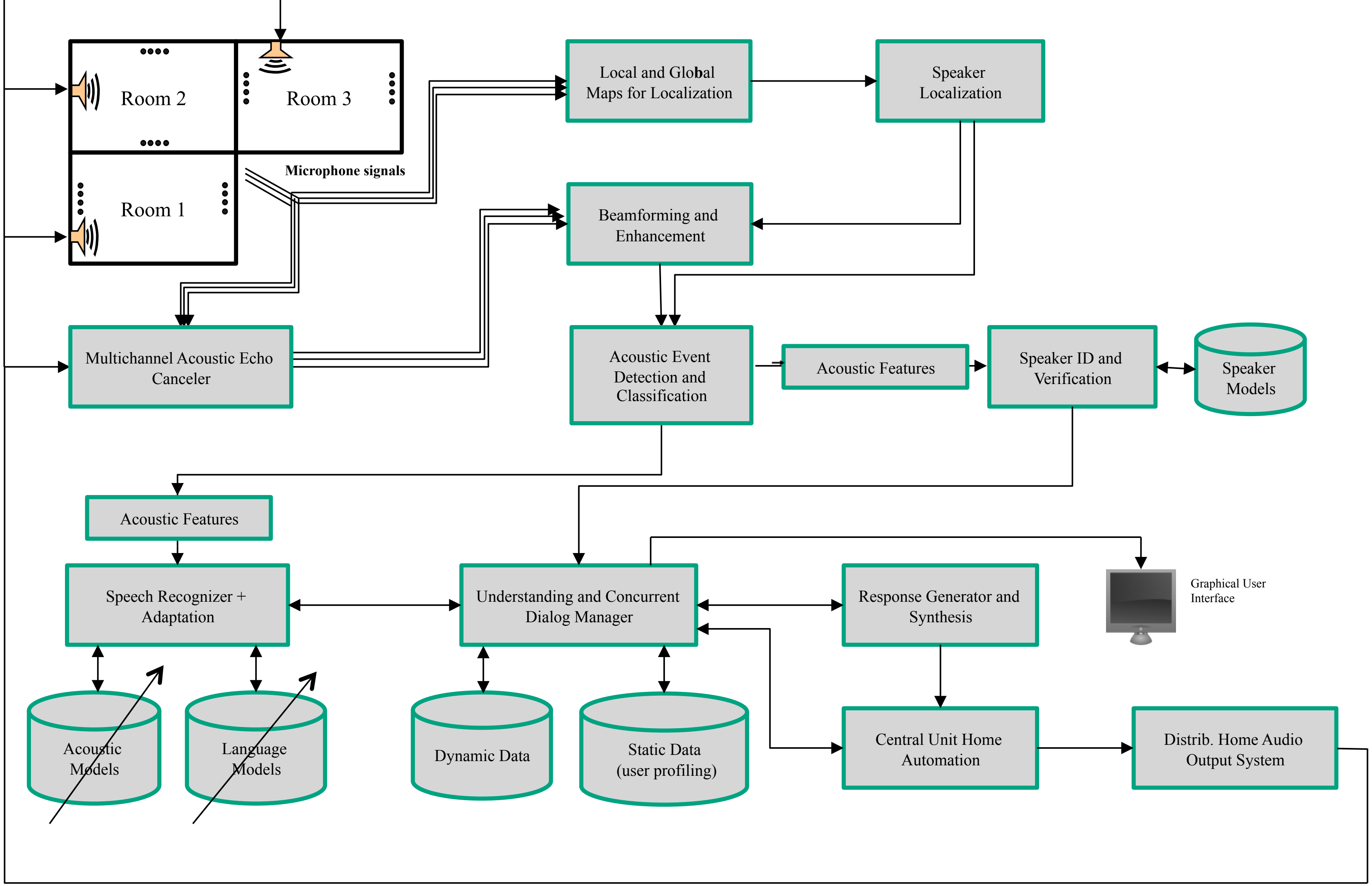


- Commands spoken in a noisy living room
- Recordings made using a binaural manikin
- Different SNR (-6dB…9dB)

Noisy **Processed**

Distant-speech interaction: DICIT project

Targeted DIRHA prototype: architecture under study



DICIT- scenario and addressed problems

- Acoustic event detection - classification of the nature of the active source
- Location of active sources
- Head orientation estimation
- Selective acquisition of a speech utterance and its enhancement
- Cancelling what is *known*
- Possible separation of simultaneously active sources
- Distant-talking speaker identification/verification
- Robust speech recognition and understanding
- Multi-modal spoken dialogue management
- Feedback to the user

For more details see the web site http://dirha.fbk.eu

DICIT Distant-talking Interfaces for Control of Interactive TV

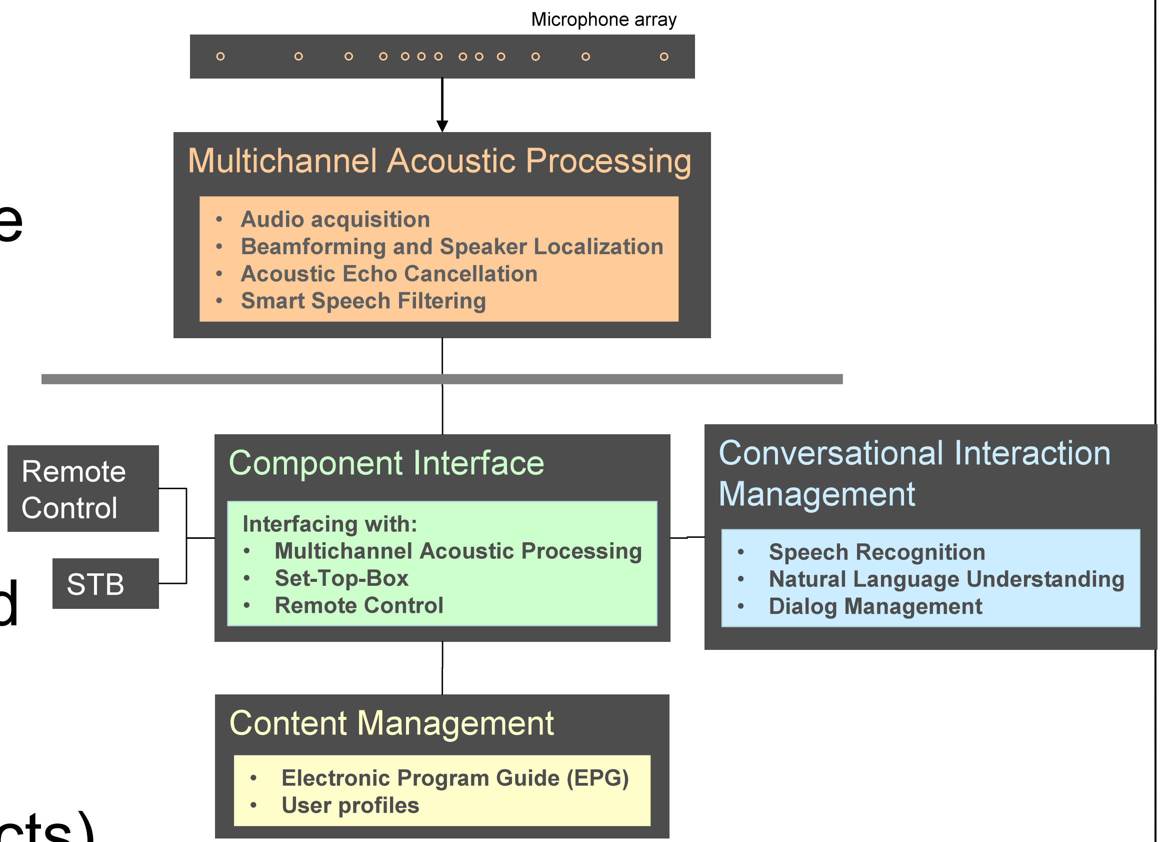
Example of interaction with one of the DICIT prototypes

- Command-and-control task
- Mono-AEC
- Multi-step GCF-based location algorithm for multiple speakers/ noise sources
- Limited inter-action area
- English and Italian languages
- Speaker ID
- Speaker inde-pendent FBK-irst speech recognizer
- Use of real STB
- Video-clip recor-ded at **ICT 2008**

Maurizio Omologo – omologo@fbk.eu

The DICIT final prototype

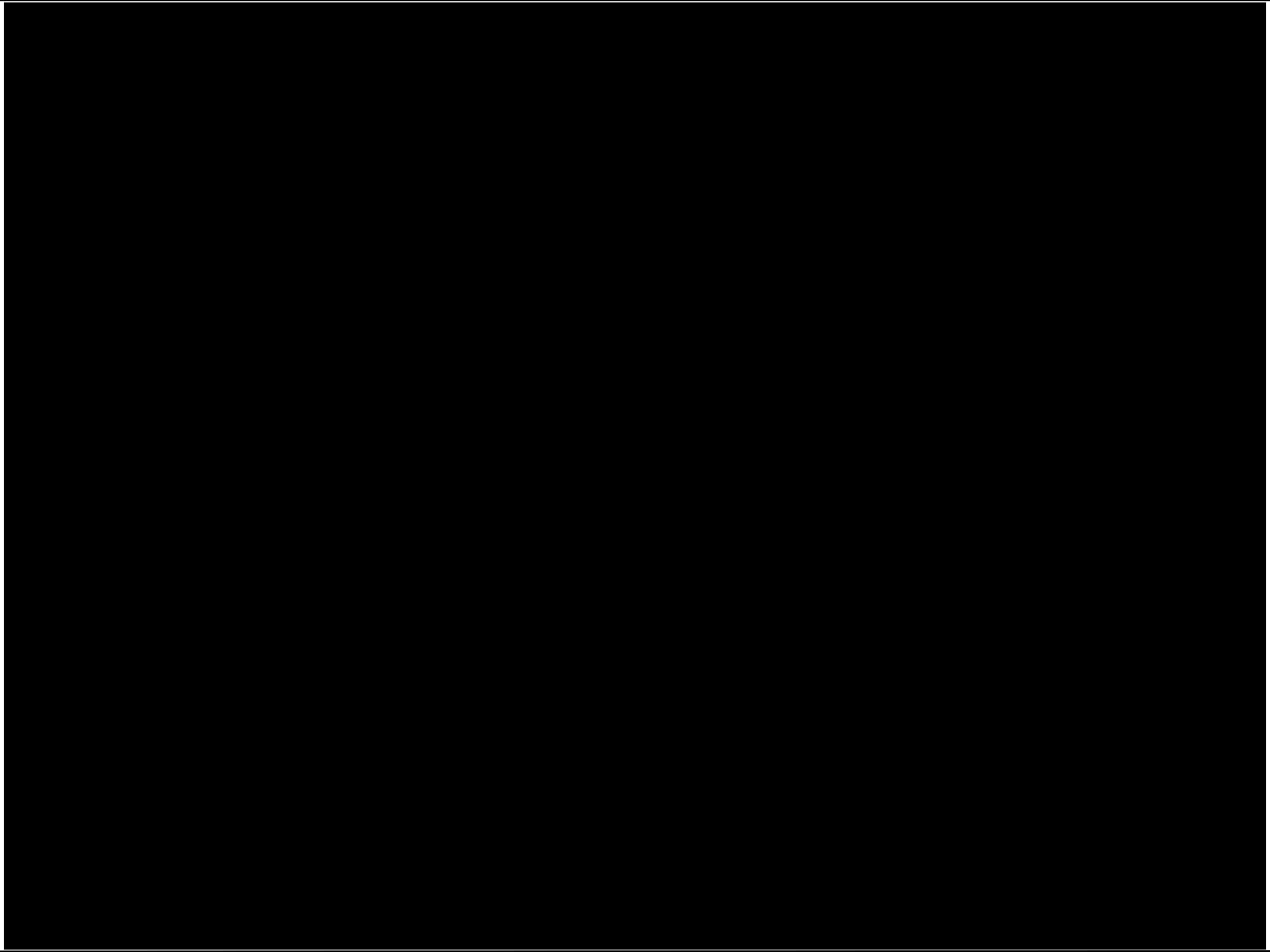
- Multi-channel acoustic processing for acoustic scene analysis
- HMM-based ASR
- Three languages (D, E, I)
- Dialogue management based on natural language understanding
- Interfaced to a real STB
- Multi-modal (speech + remote control)
- Output both graphical and as synthetic voice
- Evaluation performed in seven sites (on 172 subjects)



Block diagram of DICIT system: the upper part runs on a PC; the lower part runs on a second PC interfaced to STB, TV, and remote control.

➢ VIDEO

The DICIT final prototype



Maurizio Omologo – omologo@fbk.eu

The DICIT final prototype: example of interaction in noisy conditions

DICIT
Distant-talking Interfaces for Control of Interactive TV

- This video was recorded at IFA2009-TecWatch, in a real context characterized by very adverse noisy conditions
- During the prototype evaluation campaign, in general the system had to deal with quite more complex queries
- Other examples can be found in the website

Maurizio Omologo – omologo@fbk.eu

Discussion and Conclusions

- Distant-speech interaction involves several processing steps from microphone signal to speaker location, to enhancement, understanding and dialogue management
- Under unconstrained conditions it is a very challenging task
- Robustness can be improved in a context of distributed microphone network combined with acoustic scene analysis
- Current efforts under DIRHA towards a flexible framework for unconstrained interaction in the domestic environment using very low cost microphones.
- Multi-microphone processing technologies recalled today in this lecture are also exploited in multi-modal applications, for instance for human activity analysis based on audio-visual cues:
  - Based on past experience, robustness of each component is a fundamental issue
  - Complementarity is another principle; in some cases, audio-based information can be more accurate than video-based one (and viceversa).
  - Confidence measure across different modalities becomes a third important issue

Thanks for your kind attention!

References

- H. Kuttruff, *"Room Acoustics",* Elsevier Applied Science, (3rd edition) 1991.
- J. Blauert, *"Spatial Hearing",* MIT Press, (Revised Edition) 1997.
- D. Johnson, D. Dudgeon, *"Array Signal Processing – Concepts and Techniques",* Prentice Hall, 1993.
- M. Omologo, P. Svaizer, R. De Mori, *"Acoustic transduction",* ch. 2 of *"Spoken dialogues with computers",* R. De Mori ed., Academic Press, 1998.
- M. Brandstein and D. Ward eds, *"Microphone Arrays"*, Springer Verlag, 2001.
- Y. Huang, J. Benesty, G.W. Elko, *"Microphone arrays for video camera steering",* ch. 11 of *"Acoustic signal processing for telecommunication",* S.L. Gay and J. Benesty eds., Kluwer, 2000.
- X. Huang, A. Acero, and H.W. Hon, "*Spoken language processing: a guide to theory, algorithm, and system development*", Prentice Hall, Upper Saddle River, NJ, USA, 2001
- Y. Huang, J. Benesty, "*Audio Signal Processing for Next-Generation Multimedia Communication Systems*", Chapters 8 and 9, Kluwer 2004
- M. Wölfel and J. McDonough, "*Distant Speech Recognition*", John Wiley and Sons, 2009.

References

• A. Hyvärinen, J. Karhunen and E. Oja, “*Independent Component Analysis*”, John Wiley and Sons, 2001.

• Y. Wang, L. Deng and A. Acero, “Spoken Language Understanding - An Introduction to the Statistical Framework”, IEEE Signal Processing Magazine, vol. 22(5), 2005.

• H. Buchner, R. Aichner and W. Kellermann, “TRINICON-based blind system identification with application to multiple-source localization and separation”, In S. Makino, T.-W. Lee and S. Sawada (eds.), Blind Speech Separation, Springer-Verlag, Berlin/Heidelberg, 2007.

• I. Lee, T. Kim and T. Lee, “Independent Vector Analysis for Convolutive Blind Speech Separation”, In S. Makino, T.-W. Lee and S. Sawada (eds.), Blind Speech Separation, Springer-Verlag, Berlin/Heidelberg, 2007.

• H. Sawada, S. Araki and S. Makino, “Frequency-domain blind source separation”, In S. Makino, T.-W. Lee and S. Sawada (eds.), Blind Speech Separation, Springer-Verlag, Berlin/Heidelberg, 2007.

• A. Waibel, R. Stiefelhagen Eds., “Computers in the Human Interaction Loop”, Human Computer Interaction series, Springer Verlag, 2009.

• K. Jokinen and M. McTear,“Spoken Dialogue Systems”, Morgan&Claypool Publ.2010.

• W. Minker, G. Lee, S. Nakamura and J. Mariani “Spoken Dialogue Systems Technology and Design”, Springer, 2010.

References

- C.H. Knapp, G.C. Carter, *"The generalized correlation method for estimation of time delay",* IEEE Trans. on ASSP, vol. 24, pp. 320-327, 1976.
- J.B. Allen, D.A. Berkley, "*Image method for efficiently simulating small-room acoustics*", JASA, vol. 65, pp. 943-950, 1979.
- J. Smith, J. Abel, *"Closed-form least-squares source location estimation from range-difference measurements", IEEE Trans. on ASSP, vol. 35, pp. 1661-1669,* 1987.
- V. M. Alvarado, "*Talker localization and optimal placement of microphones with a linear microphone array using stochastic region contraction*", PhD Thesis, Brown University, 1990.
- M. Omologo and P. Svaizer, "*Use of the Cross-power Spectrum Phase in Acoustic Event Localization*", ITC-irst Tecnical Report #9303-13, March 1993.
- J. L. Flanagan, A. Surendran, E. Jan, "*Spatially selective sound capture for speech and audio processing*", Speech Commnication, vol. 13, pp. 207-222, 1993.
- M. Omologo, P. Svaizer, *"Acoustic event location using a Crosspower-Spectrum Phase based technique",* Proc. of IEEE ICASSP 1994, pp.273-276.

References

• Y. Chan, K. Ho, "*A simple and efficient estimator for hyperbolic location*", IEEE Trans. Signal Processing, vol.42, pp. 1905-1915, 1994.

• M.S. Brandstein, J.E. Adcock, H.F. Silverman, *"A practical time delay estimator for localizing speech sources with a microphone array",* Comp. Speech Language, vol. 9, pp.153-169, 1995.

• Y. Suzuki, F. Asano, H.Y. Kim, T. Sone, "*An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses*", JASA, vol. 97, pp. 1119-1123, 1995.

• D.V. Rabinkin, R.J. Renomeron, J.C. French, J.L. Flanagan,"*A DSP implementation of source location using microphone arrays*", Proc. SPIE, 1996.

• E.E. Jan and J.L. Flanagan, "*Sound source localization in reverberant environments using an outlier elimination algorithm*", Proc. of ICSLP 1996.

• B. Champagne, S. Bédard, A. Stéphenne, *"Performance of time delay estimation in the presence of room reverberation",* IEEE Trans. on SAP, vol. 4, pp. 148-152, 1996.

• M. Omologo, P. Svaizer, "*Use of the crosspower-spectrum phase in acoustic event location*", IEEE Trans. on SAP, vol. 5, pp. 288-292, 1997.

References

• H. Wang, P. Chu, "*Voice source location for automatic camera pointing system in videoconferencing*", in Proc. of IEEE ICASSP 1997, pp. 187-190.

• P. Svaizer, M. Matassoni, M. Omologo, "*Acoustic source location in a three-dimensional space using crosspower spectrum phase*", in Proc. of IEEE ICASSP 1997, pp. 231-234.

• M.S. Brandstein, H.F. Silverman, "*A practical methodology for speech source location with microphone arrays*", Comp. Speech Language, vol. 11, pp. 91- 126, 1997.

• M.S. Brandstein, J.E. Adcock, H.F. Silverman, "*A closed-form location estimator for use with room environment microphone arrays*", IEEE Trans. SAP, vol. 5, pp. 45-50,1997

• P.G. Georgiou et al., " Alpha-stable modeling of noise and robust time delay estimation in the presenc of impulsive noise", IEEE Trans. on Multimedia, vol. 1, n. 3, pp. 291-301, 1999.

• J. Benesty, *"Adaptive eigenvalue decomposition algorithm for passive acoustic source location",* JASA, vol.107, pp. 384-391, 2000.

• T. Nishiura, T. Yamada, S. Nakamura, K. Shikano , "*Location of multiple sound source based on a CSP analysis with a microphone array* ", Proc. of IEEE ICASSP 2000, pp. 1053-1056.

References

- J. Vermaak and A. Blake, “*Nonlinear filtering for Speaker tracking noisy and reverberant environments*”, Proc. of ICASSP 2001.
- J. Vermaak et al., “*Sequential monte carlo fusion of sound and vision for speaker tracking*”, Int. Conf. on Computer Vision, 2001.
- D. Zotkin et al., “*Multimodal 3-D tracking and event detection via the particle filter*”, IEEE Workshop on Detection and Recognition of Event in Video, 2001
- N. Strobel, S. Spors, and R. Rabenstein, “*Joint audio-video object localization and tracking*“, IEEE Signal Processing Magazine, vol. 18,Jan. 2001.
- Y. Huang et al., “*Real-time passive source localization: A practical Linear-correction least-squares approach*”, vol. 9, n. 8 November 2001.
- D.B. Ward and R.C. Williamson, “*Particle filter beamforming for acoustic source localization in a reverberant environment*”, Proc. of ICASSP 2002.
- M. Matassoni, M. Omologo, D. Giuliani and P. Svaizer, “*Hidden Markov model training with contaminated speech material for distant-talking speech recognition*”, Computer Speech & Language, vol. 16(2),(2002).
- J. Chen, J. Benesty, Y. Huang, *“Robust time delay estimation exploiting redundancy among multiple microphones”,* IEEE Trans. on SAP, vol. 11, 2003.
- T.G. Dvorkind and S. Gannot, “*Speaker Localization exploiting spatial-temporal information*”, IEEE Workshop on Ac. Echo and Noise control, September 2003.

References

- L. Armani, M. Matassoni, M. Omologo, P. Svaizer, "*Use of a CSP-based voice detector for distant-talking ASR*", Proc. of EUROSPEECH, Geneva, Switzerland, September 2003.
- E. Lehmann et al., “*Experimental comparison of particle filtering algorithms for acoustic source localization in reverberant room*”, Proc. of ICASSP 2003.
- D. B. Ward et al., “*Particle filtering algorithms for tracking an acoustic source in a reverberant environment*”, IEEE Trans. on SAP, vol. 11, n.6, pp. 826-836, 2003.
- S. Doclo and M. Moonen, “*Robust Adaptive Time Delay Estimation for Speaker Localization in Noisy and reverberant Acoustic Environments”*, EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1110-1124, 2003.
- H. Asoh et al., “*An application of a particle filter to bayesian multiple sound source tracking with audio and video information fusion*”, in Proc. Fusion, 2004, pp. 805-812.
- J. Benesty et al., “*Time-delay estimation via Linear Interpolation and cross-correlation*”, IEEE Trans. on Speech and Audio Processing, vol. 12, n. 5, 2004.
- A. Brutti, M. Omologo, P. Svaizer, “*Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays*”, Proc. of Interspeech 2005.

References

• H. Buchner et al.,"*Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering*", Proc. of ICASSP, vol. III, pp. 97-100, 2005

• H. Teutsch, W. Kellermann, " *EB-ESPRIT: 2D localization of multiple wideband acoustic sources using eigen-beams*" ,Proc. of ICASSP 2005

• L. Brayda, C. Bertotti, L. Cristoforetti, M. Omologo, P. Svaizer, "*Modifications on NIST MarkIII array to improve coherence properties among input signals*" 118th AES Convention, 2005.

• D. Macho et al., "*Automatic Speech Activity Detection, Source Localization, and Speech Recognition on the CHIL Seminar Corpus*", Proceedings of ICME, 2005.

• M. Omologo et al., "*Speaker Localization in CHIL lectures: Evaluation Criteria and Results*", MLMI'05, Springer Lecture Notes in Computer Science vol. 3869, 2006.

• G. Lathoud and J.M. Odobez, "*Short-term spatio-temporal clustering applied to multiple moving speakers*", IEEE Trans. on Audio, Speech and Language Processing, vol. 15(5), July 2007.

• E. Lehman and A. Johansson, "*Particle filter with integrated voice activity detection for acoustic source tracking*", EURASIP Journal on Applied Signal Processing, vol. 2007.

References

- C. Zieger, M. Omologo, "*Acoustic Event Classification Using a Distributed Microphone Network with a GMM/SVM Combined Algorithm*", Interspeech 2008.
- E. Zwyssig, M. Lincoln and S. Renals (2010), "*A Digital Microphone Array for Distant Speech Recognition*", ICASSP-2010.
- X. Wu, T. Ren and L. Liu, "*Sound source localization based on directivity of MEMS microphones (Vol. 3)*", 7th International Conference on Solid-State and Integrated Circuits Technology, 2004.
- A. Brutti, M. Omologo and P. Svaizer, "*Multiple Source Localization Based on Acoustic Map De-Emphasis*", EURASIP Journal on Audio, Speech, and Music Processing, vol. 2010.
- H. Christensen, J. Barker, N. Ma and P. Green. "*The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments*". In Proc. Interspeech 2010.
- C. Bartsch, A. Volgenandt, T. Rohdenburg, and J. Bitzer, "*Evaluation of different microphone arrays and localization algorithms in the context of ambient assisted living*", IWAENC 2010.

Most popular BSS approaches

	PROS	CONS
Time-domain [20][23]	+ Theoretically optimal	- Low convergence - High risk of divergence or local minima - Hard to generalize to the underdetermined case
Frequency-domain [25]	+ Computationally efficient + Reduces risk of local minima + Extendable to the underdetermined case	- Permutation and scaling ambiguity - Statistically biased: few data observed in each frequency
Multivariate (e.g. IVA [21])	+ A trade-off between time-domain/frequency-domain + No permutation ambiguity	- Low convergence - High risk of divergence or local minima - Hard to generalize to the underdetermined case
Sparseness based [24](e.g. DUET)	+ Computationally efficient + Implicit models the underdetermined case	- In echoic environments do not work as good as in anechoic environment (audible distortions)

References about BSS and source extraction (1)

[1] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariancemodel. *Audio, Speech and Language Processing, IEEE Transactions on, Special Issue on Processing Reverberant Speech*, 2010.

[2] A. Brutti and F. Nesta. "Multiple source tracking by sequential posterior kernel density estimation through GSCT" In *Proceedings of the European Signal Processing Conference*, Barcelona, Spain, 2011.

[3] R. DeMori. *Spoken Dialogues with Computers*. Academic Press, London, 1998.Chapter 2.

[4] T. Gustaffson, B. D. Rao, and M. Trivedi. Source localization in reverberant environments: Modeling and statistical analysis. *Speech and Audio Processing, IEEE Transactions on*, 11(6):791–803, Nov. 2003.

[5] F. Nesta and A. Brutti. Self-clustering non-euclidean kernels for improving the estimation of multidimensional TDOA of multiple sources. In *Workshop on Handsfree Speech Communication and Microphone Arrays*, 2011.

[6] F. Nesta and M. Omologo. Generalized State Coherence Transform for multidimensional TDOA estimation of multiple sources. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2011.

[7] F. Nesta and M. Omologo. Enhanced multidimensional spatial functions for unambiguous localization of multiple sparse acoustic sources. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Kyoto, Japan, 2012.

[8] P. Svaizer, A. Brutti, and M. Omologo. Use of reflected wavefronts for acoustic source localization with a line array. In *Workshop on Hands-free Speech Communication and Microphone Arrays*, 2011.

[9] A. Brutti, F. Nesta " Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs", to appear in Elsevier Computer Speech and Language

[10] F. Nesta, M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments", to appear in Elsevier Computer Speech and Language 2012

[11] J. Barker, E. Vincent, N.Ma, H. Christensen, and P. Green. "The PASCAL CHiME speech separation and recognition challenge." to appear in Elsevier Computer Speech and Language 2012

[12] F. Nesta and M.Matassoni. "Robust automatic speech recognition through on-line semi-blind source extraction." In *Proceedings of CHIME*, Florence, Italy, 2011.

[13] F. Nesta, M. Matassoni, and H. Maganti. "Real-time prototype for integration of blind source extraction and robust automatic speech recognition." In *Proceedings of Interspeech*, 2011.

[14] F. Nesta and M. Omologo. "Convolutive underdetermined source separation through weighted interleaved ICA and spatio-temporal correlation." In *Proceedings LVA/ICA*, Mar 2012.

References about BSS and source extraction (2)

[15] F. Nesta, T. Wada, and B.-H. Juang. "Batch-online semi-blind source separation applied to multi-channel acoustic echo cancellation." *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(3):583 –599, 2011.

[16] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. "Blind spatial subtraction array for speech enhancement in noisy environment." *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):650 –664,May 2009.

[17] M. Fakhry, F. Nesta "Underdetermined Source Detection and Separation Using a Normalized Multichannel Spatial Dictionary", IWAENC 2012

[18] A. Brutti, M. Omologo, P.G. Svaizer, "Oriented global coherence field for the estimation of the head orientation in smart rooms equipped with distributed microphone arrays". Eurospeech 2005, Lisboa.

[19] Shun-ichi Amari, "Natural Gradient Works Efficiently in Learning", Neural Computation feb. 1998

[20] Herbert Buchner, Robert Aichner, and Walter Kellermann. TRINICON: A versatile framework for multichannel blind signal processing. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 889–892, Montreal, Canada, May 17-21 2004.

[21] Intae Lee, Taesu Kim, and Te-Won Lee. Independent vector analysis for convolutive blind speech separation. In *Blind Speech Separation*. Springer, September 2007.

[22] C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 24, pages 320–327, 1976.

[23] Z. Koldovsk`y and P. Tichavsk`y. Time-domain blind audio source separation using advanced component clustering and reconstruction. In *Proceedings of HSCMA*, Trento, Italy, May 2008.

[24] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[25] H. Sawada, S. Araki, and S. Makino. "Frequency-domain blind source separation. In *Blind Speech Separation*." Springer, September 2007.

[26] K. Matsuoka and S. Nakashima. Minimal distortion principle for blind source separation. In *Proceedings of International Symposium on ICA and Blind Signal Separation*, San Diego, CA, USA, December 2001.