Far-Field ASR in Greek for Domestic Environment and Child-Robot-Interaction Applications

Gerasimos Potamianos [1,3] Petros Maragos [2,3]

[1] ECE Dept., University of Thessaly, Volos, Greece

[2] School of ECE, National Technical University of Athens, Greece

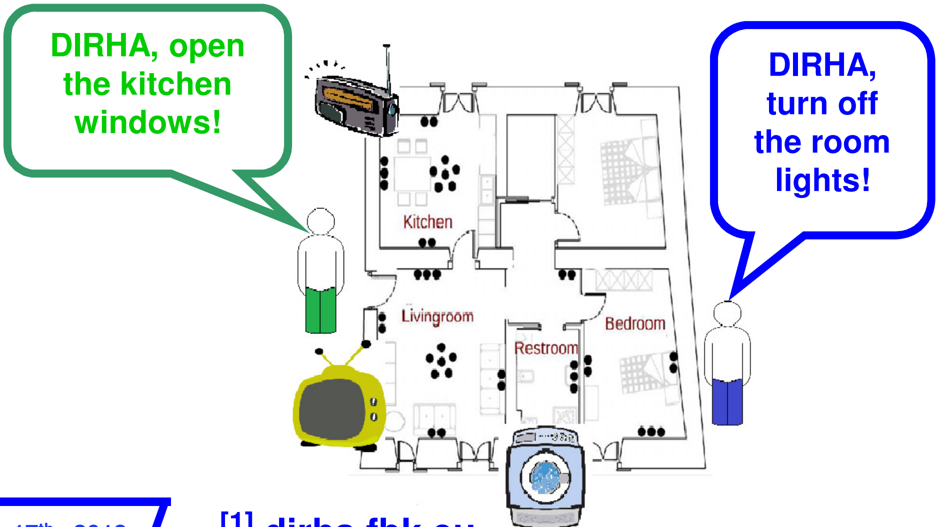
[3] Athena Research & Innovation Center, Maroussi, Greece

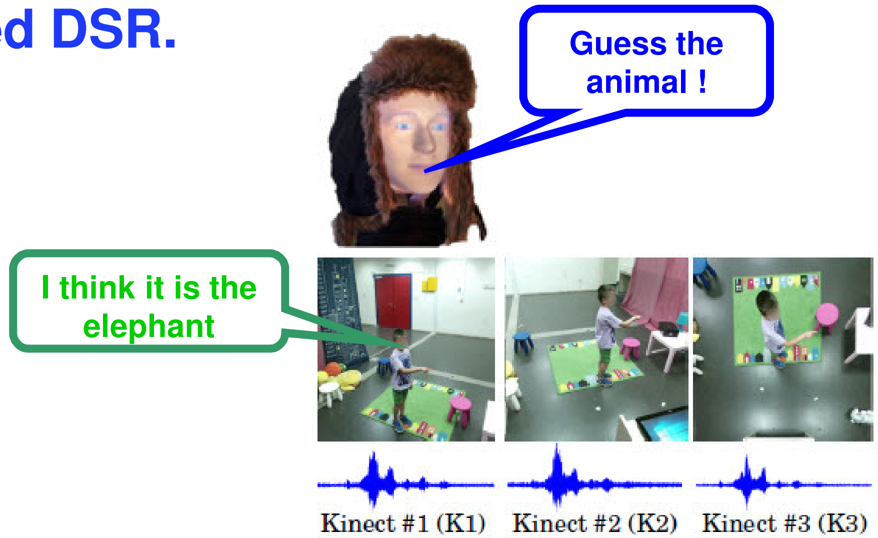
LISTEN Workshop / Summer School
July 17-19, 2018
Bonn, Germany

2018.07.17

Focus of this Presentation

- Work on two **EU projects** with **far-field multichannel ASR** components:
  - ✓ **DIRHA** (2012 -14): *Distant-speech Interaction for Robust Home Applications* [1]
  - ✓ **BabyRobot** (2016 -18): *Child-Robot Communication and Collaboration – Edutainment, Behavioural Modelling and Cognitive Development in Typically Developing and Autistic Spectrum Children* [2]
- Our work focus lies on ASR in **Greek** for the specific **project scenarios**.
  - ✓ **Always-listening, command-based DSR.**



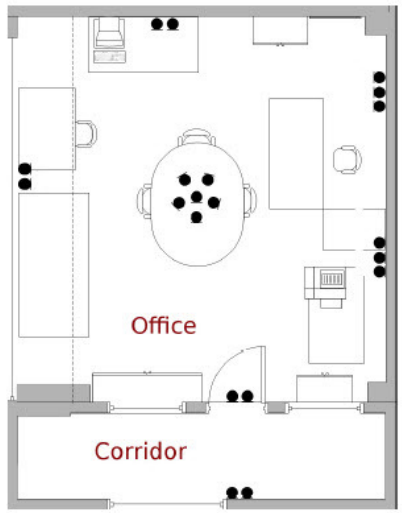


[1] dirha.fbk.eu

[2] babyrobot.eu

Part I: DIRHA Project

- **Motivation.**
- **System components at a glance.**
- **DSR system.**
- **Corpora and results.**
- **A module at detail: Multi-room SAD.**



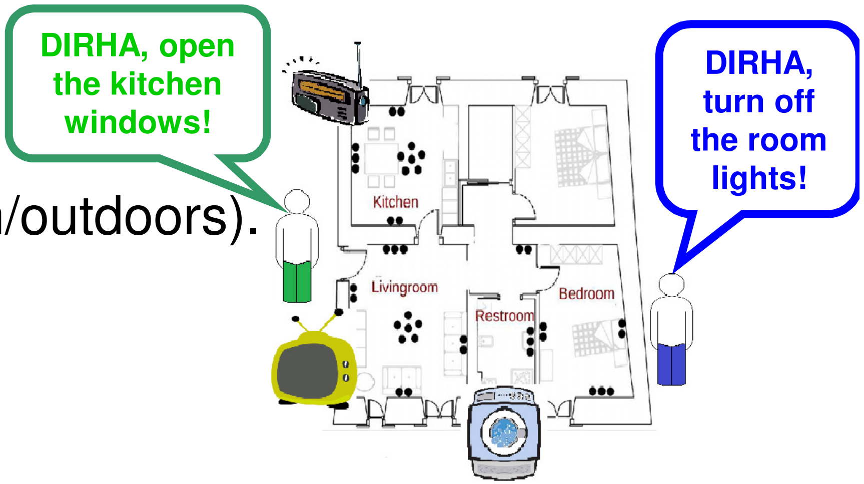
-- **DIRHA:** *Distant-speech Interaction for Robust Home Applications* (**dirha.fbk.eu**)

-- Rodomagoulakis *et al.,* "Room-localized command recognition in multi-room, multi-microphone environments, *CSL'17*.

-- Giannoulis *et al.,* "Multi-room speech activity detection using a distributed microphone network in domestic environments", *Eusipco'15*.

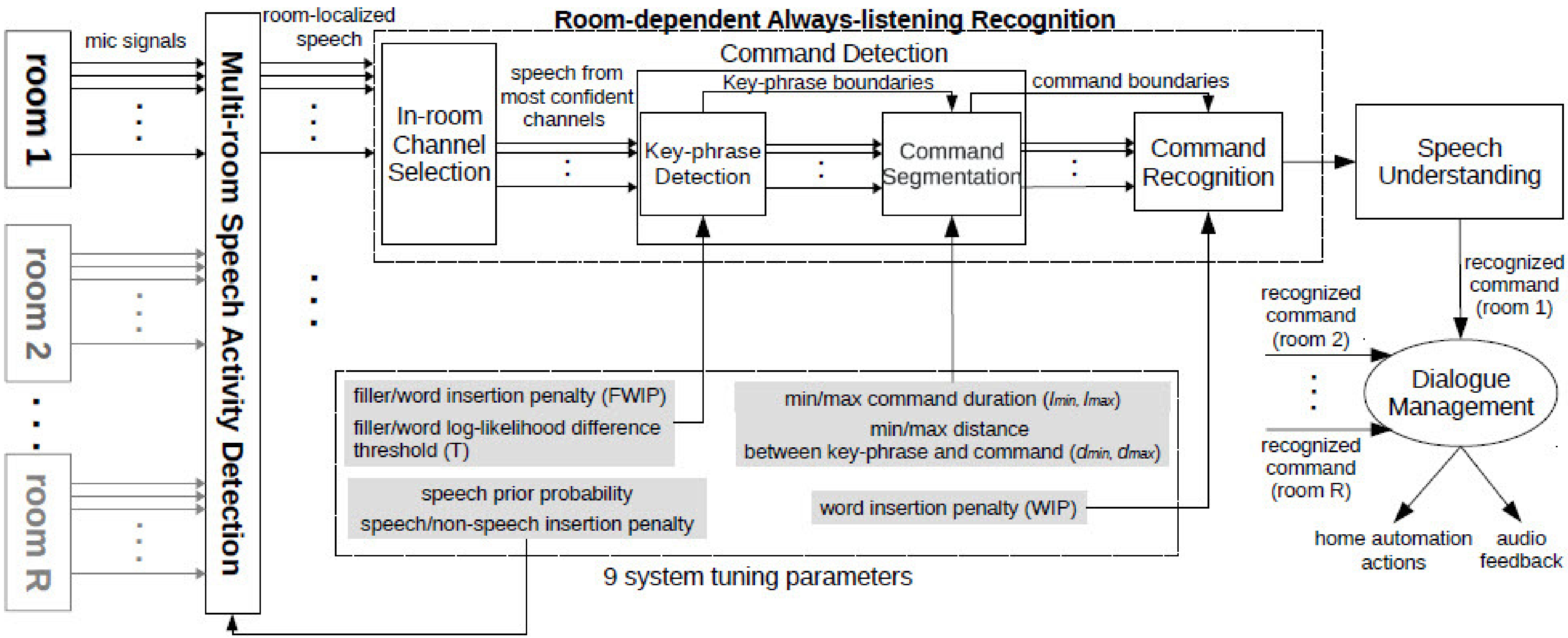
Motivation

- Towards **voice-enabled smart-homes** …
  - ✓ Natural, seamless **control** of **domestic devices** (doors, windows, ...).
  - ✓ Improved **safety** and **comfort** (disabled users, ambient assistive living).
  - ✓ Focus of many recent **projects** (SweetHome, **DIRHA**, ...).
  - ✓ “Holy grail”: **always-listening**, **far-field**, **robust** operation.
- **Difficult goal** in practice, due to **challenging domestic acoustic scene**:
  - ✓ Signal attenuation (**low SNR**).
  - ✓ Signal reflections (**reverberation**).
  - ✓ Multiple **speech** & **noise** sources (in/outdoors).
  - ✓ Possible speech & noise **overlap**.
  - ✓ **Inter-room interference**.
- Promising **mitigation**:
  - ✓ **Multi-channel** approaches (**microphone-array** sensors).



System Block Diagram

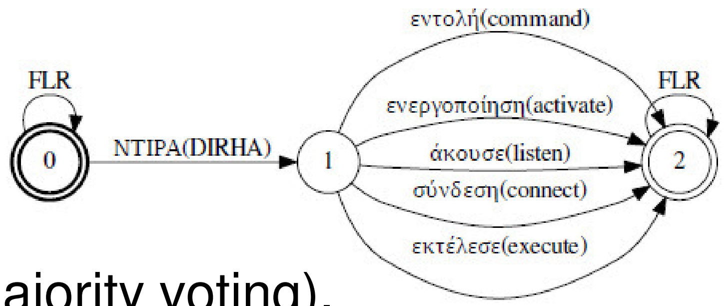
- **Parallel DSR pipelines**, per room, for **multi-room homes** [3].
  - ✓ Driven by "**room-dependent**" **SAD**.
- Room-level pipeline **components:**
  - ✓ Channel **selection**; **key-phrase** detection; command **segmentation**; command **recognition**.



[3] Rodomagoulakis *et al.*, "Room-localized command recognition in multi-room, multi-microphone environments, *CSL*'17.

System Modules (after SAD)

- **In-room channel selection**:
  - ✓ Based on **envelop variance (EV)** measure.
  - ✓ Up to **top-4** microphones selected for decision fusion in next modules.
- **Key-phrase detection:**
  - ✓ Based on classical **keyword-filler KWS** approach.
  - ✓ Traditional **MFCC+derivs**. front-end, **GMM-HMM** acoustic modeling.
  - ✓ Filler model: **24** states, **32** mix/state.
  - ✓ Key-phrases: **12** in total.
  - ✓ **Training:** discussed in ASR module.
  - ✓ **Testing:** decision fusion over 4 mics (majority voting).



- **Command segmentation:**
  - ✓ Based on in-room **SAD segments** and heuristics of **duration / distance** following key-phrase detection.

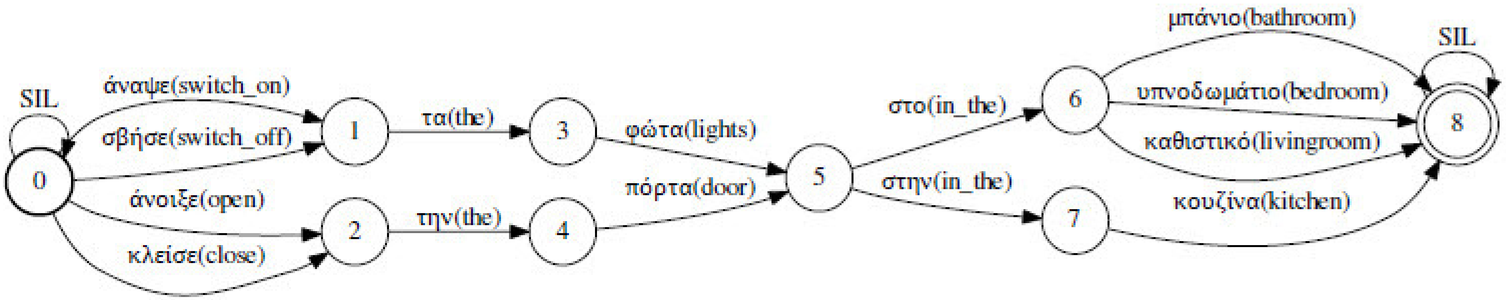
listen WORKSHOP

DSR Module (I) – Training

- **Close-talk model training ("CLEAN")**:
  - ✓ Traditional **MFCC+derivs**. front-end, **GMM-HMM** acoustic modeling.
  - ✓ About 8k CD triphones, with 16 mix/state.
  - ✓ Corpus: "**Logotypographia**" (Greek set, 125 spk, **72 hrs**, 50k wds).
  - ✓ Close-talking part of it used (75 spk, **22.6 hrs**).
- **Far-field models ("REVERB"):**
  - ✓ Trained on **artificially contaminated** Logotypographia data with **RIRs** (T_{60} = 0.7 s), available from the DIRHA project, plus **white Gaussian noise**, simulating **far-field** conditions.
- **Further robustness:**
  - ✓ Supervised **MLLR adaptation** on **in-domain** dev. data.
  - ✓ Models **per microphone** (not per speaker).

DSR Module (II) – Testing

- **Closed-grammar decoding**:
  - ✓ **180** home-automation commands.



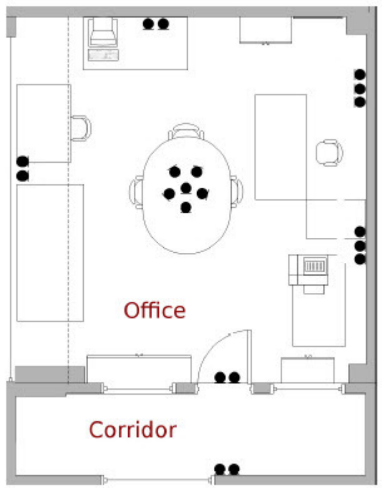
- **Multi-microphone decision fusion:**
  - ✓ **Best-EV** microphone signal decoded.
  - ✓ **N-best** results obtained (N = **3**).
  - ✓ Rescored by **top-3** microphones (forced alignment).
  - ✓ Obtained scores **averaged** and **max** obtained.
- **Signal fusion** also considered (**6** channels used):
  - ✓ Using **MVDR beamforming.**
  - ✓ **Wiener** post-filter with weights estimated by **MMSE**.

Databases

- **Three corpora**:
  - ✓ **Simulated** and **real** data recordings.
  - ✓ 2 **environments** (**DIRHA** apt. & **Athena-RC** office).

data characteristics	databases		
	DIRHA-sim	DIRHA-real	ATHENA-real
rooms (#)	4	4	2
microphones (#)	40	40	20
subjects (#)	20	5	20
background noises (#)	10	not transcribed	4
non-speech events (#)	73	not transcribed	15
total speech (min)	37	18	72
unique commands (#)	99	59	172
activation phrases (#)	12	12	12
avg SNR (dB)	13	15	9
avg T_{60} (sec)	0.72	0.72	0.50
close-talk mic available	no	no	yes

DIRHA apartment @ FBK ~ 50m²



Athena-RC office ~ 35m²

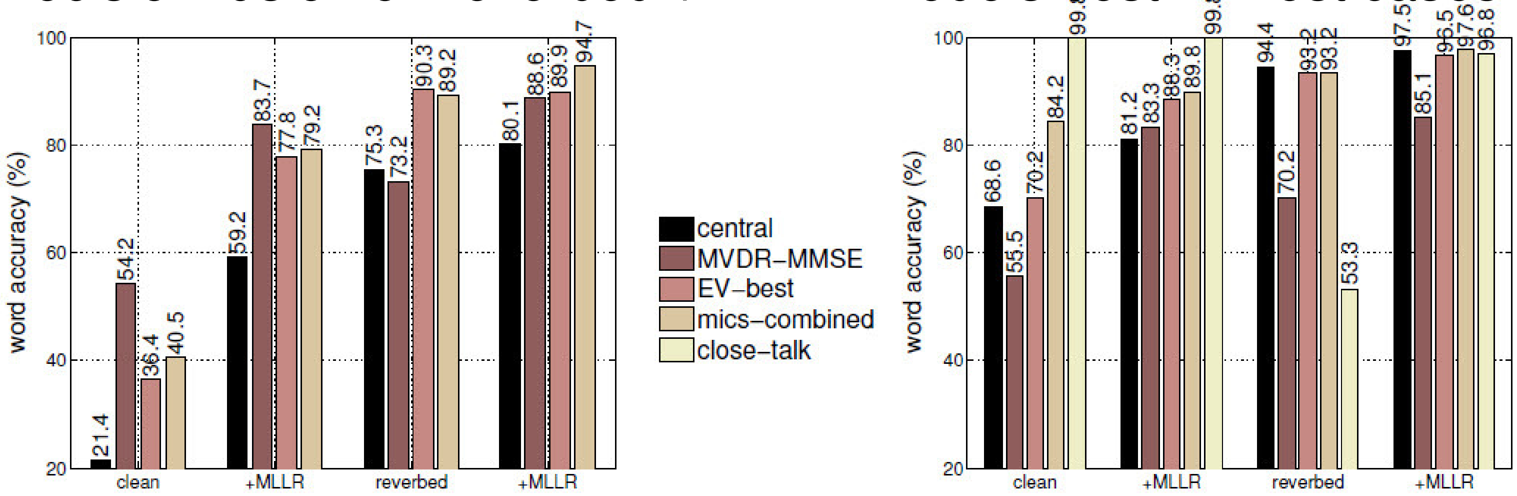
-- Cristoforetti *et al.*, “The DIRHA simulated corpus”, *LREC’14*.
-- Matassoni *et al.*, “The DIRHA-GRID corpus: Baseline and tools …”, *Interspeech’14*.
-- Tsiami *et al.*, “ATHENA: A Greek multi-sensory database …”, *Interspeech’14.*

Results

- **Performance of overall system in Sentence Accuracy (%):**
  - ✓ **Baseline:** clean models + MLLR, 1 mic (EV-best), separate module opt.
  - ✓ **Proposed:** reverb models + MLLR, decision fusion, joint module opt.

	Baseline	Proposed
DIRHA-sim	**29.3**	**38.7**
DIRHA-real	**45.0**	**60.0**
ATHENA-real	**59.7**	**76.6**

- **Channel selection / fusion experiment (DSR with ground truth segm.)**
  - ✓ Decision fusion on reverbed + MLLR models best in most cases.



(a) DIRHA-sim (b) ATHENA-real

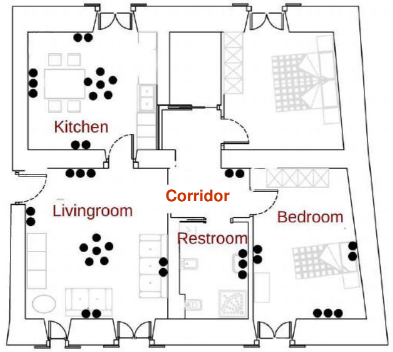
Module Detail: Room-Level SAD

- Exploit **multiple microphones** to detect **speech segments** of the **acoustic scene** inside **multi-room smart homes**.
- Perform this:
  - ✓ not only at the "**home-level**" ➔ "**room-independent**" **SAD**,
  - ✓ **but also** at the "**room-level**" ➔ "**room-dependent**" **SAD**.
- Why is "room-dependent" SAD **interesting**?
  - ✓ **Disambiguates** user input (*e.g.*, "which room lights to turn off").
  - ✓ Provides **localized** user **feedback** (loudspeaker "on" in specific room).
  - ✓ Helps **ASR** (channel selection, speaker localization, speech separation).
  - ✓ **Literature:** Focus of recent works, *e.g.* [4], [5].
- Main **ideas** [6]:
  - ✓ Proposed "room-dependent" SAD **two-stage approach**.
  - ✓ Novel acoustic **features** for room localization.

[4] Moralles-Cordovilla *et al.*, "Room localization for DSR", *Interspeech'14*.
[5] Ferroni *et al.*, "A DNN approach for VAD in multi-room domestic scenarios", *IJCNN'15*.
[6] Giannoulis *et al.*, "Multi-room speech activity detection using a distributed microphone network in domestic environments", *Eusipco'15*

SAD Formulation / Notation

- **Smart home** with:
  - ✓ R **rooms** (index: $r = 1, \ldots, R$).
  - ✓ M **mics.** (M_r mics. in room r, with $\sum_{r=1}^{R} M_r = M$).



- **Audio signal(s)**:
  - ✓ Captured by mic. m of room r : $x_m^r(t)\,,\quad m = 1, \ldots, M_r\,,\quad r = 1, \ldots, R$
  - ✓ All signals at time t : $\mathbf{x}_t = [\, x_1^1(t)\,, \ldots, x_{M_1}^1(t)\,, \ldots, x_1^R(t)\,, \ldots, x_{M_R}^R(t) \,]$
  - ✓ Observation sequence of duration T : $\mathbf{X} = [\, \mathbf{x}_1\,, \ldots, \mathbf{x}_T \,]$
- "**Room-independent**" **SAD**:

  Find state seq. $Q' = [\, q_1\,, \ldots, q_T \,]$

  that maximizes prob. $p\,(Q'|\mathbf{X})$
- "**Room-dependent**" **SAD**:

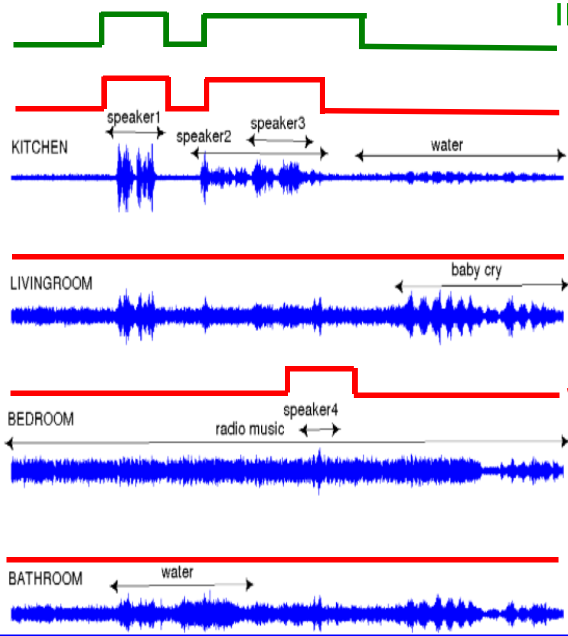
  Find seqs. $Q^r = [\, q_1^r\,, \ldots, q_T^r \,]\,,$

  for each room $r = 1, \ldots, R$,

  that maximize $p\,(Q^1, \ldots, Q^R|\mathbf{X})$

speech / non-speech states

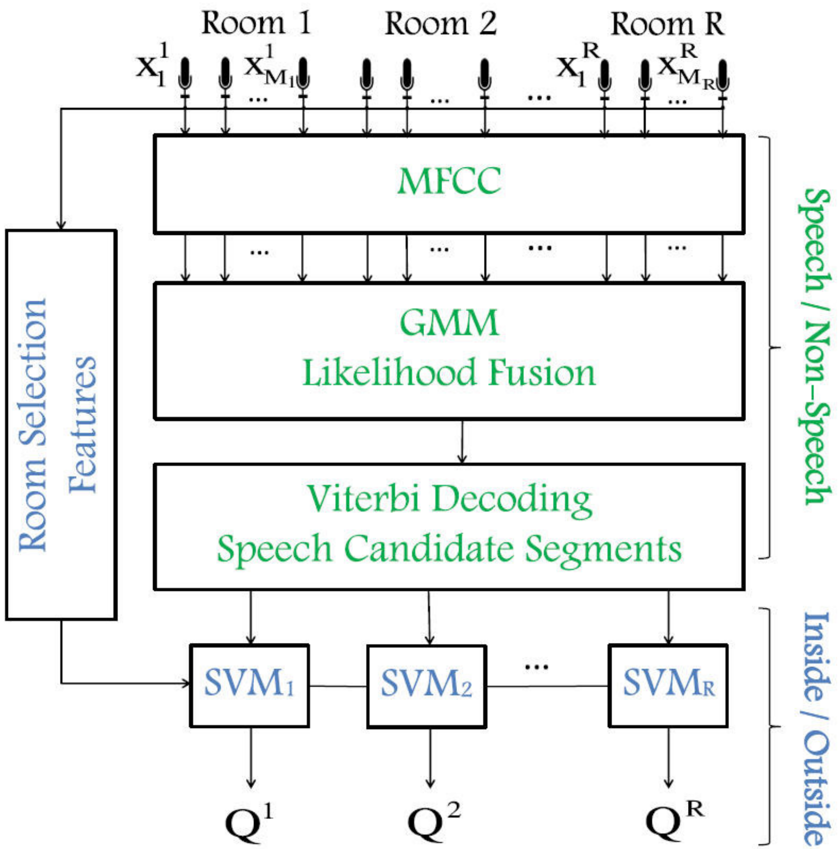
room-independent SAD



room-dependent SAD

SAD System Overview

- Proposed "**room-dependent**" **SAD** system consists of **two stages**.



1st stage: "**Room-independent**" **SAD**

- **GMMs** trained for **each** microphone.
- **Fused** by **multi-stream** framework.
- **Viterbi decoding** provides candidate speech segments to 2nd stage.

2nd stage: Inside/outside classification

- 1st-stage candidate segments get **classified** as **inside** or **outside** each **room** by room-specific **SVMs**.
- Based on **room selection features**.
- Output yields "**room-dependent**" **SAD**.

SAD – 1st Stage

Uses **all mics.** to yield "**room-independent**" speech/non-speech segmentation.

- Train a separate **two-class GMM** (speech/non-speech) **for each** microphone m of room *r*, using an **MFCC front-end**: $\lambda_m^r \,, \;\; m = 1, \ldots, M_r \,, \;\; r = 1, \ldots, R \,.$
- **Fuse** all GMM **log-likelihoods** in **multi-stream** style [7]:

$$\mathcal{L}(\, q_t \,|\, \mathbf{x}_t \,) \;=\; \sum_{r=1}^{R} \sum_{m=1}^{M_r} \log\, p \;(\, q_t \,|\, x_m^r(t) \,; \lambda_m^r \,)$$

- Use these in **Viterbi decoding** to provide the most likely **speech** / **non-speech sequence**, $Q' = [\, q_1 \,, \ldots , q_T \,]$
- State-change **penalty** in Viterbi decoding provides **smooth segmentation**.
- Resulting speech segments, (t_s, t_e) , are **passed to the 2nd stage**, to be **assigned** to room(s).

  speech segment
  start & **end times**

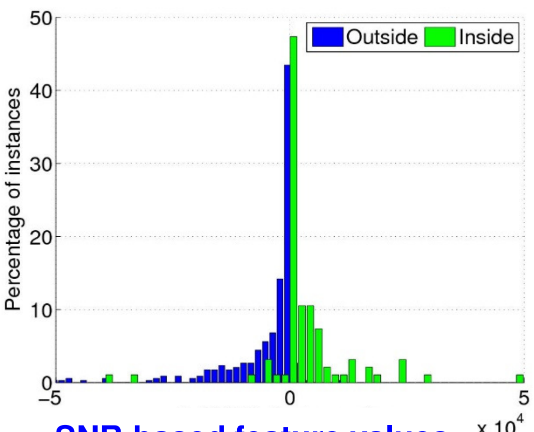
- **Implementation details**: **39**-dim MFCCs (+derivs.), **25** ms frames @ **100** Hz (10 ms window shift), **32**-mixture GMMs with diagonal covariances.

[7] Giannoulis *et al.*, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," *EUSIPCO'14*.

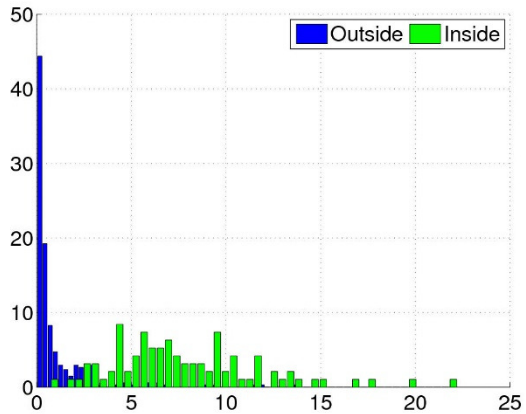
SAD – 2nd Stage: Overview

- **Main idea:**
  - ✓ For each "room-independent" **speech segment** (from 1st stage);
    - ✓ For each **room**;
      - ✓ Compute "**room-selection features**" of the segment;
        - ✓ To discriminate if it originates from **inside** vs. **outside** room.
- Need **discriminative** features. Note that "**outside**" vs. "**inside**" speech has:
  - ✓ **Lower energy** ➔ use **SNR**-based measurements.
  - ✓ **Higher reverberation** ➔ use signal **correlation**, **envelope variance**.
- We employ **three features** – their **histograms** show **room-discriminability**:

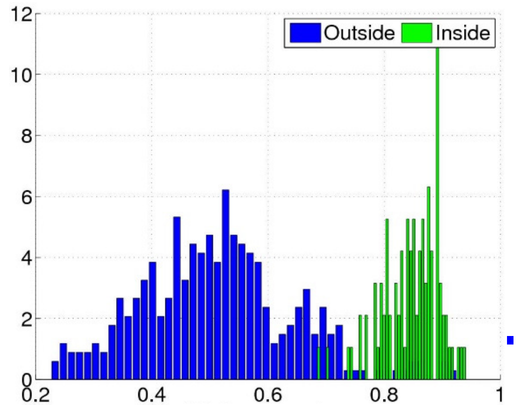
feature histograms computed on specific room of database (see later)



SNR-based feature values



correlation-based feature



envelope-variance

SAD 2nd St.: Fusion/Classification

- **Fusion** of all **features** across **rooms**:
  - ✓ For candidate **speech segment**, (t_s, t_e) ,
  - ✓ for **each room**, $r \in \{1, \ldots, R\}$, **concatenate** the three features:

$$\theta^r(t_s, t_e) = [\sigma^r(t_s, t_e), C^r(t_s, t_e), EV^r(t_s, t_e)]$$

room-specific feature vector

SNR-based feature

correlation-based

envelope variance-based

  - ✓ then, **concatenate** them over **all rooms**:

$$\theta(t_s, t_e) = [\theta^1(t_s, t_e), \ldots, \theta^R(t_s, t_e)]$$

final 3*R*-dim feature vector

3-dim room-specific features

- **Classification** as "**room-inside**" or "**room-outside**" segment:
  - ✓ Use **room-specific**, **2-class SVMs**, trained on **3*R***-dim feature vector.
  - ✓ **Score** segment by each room-specific SVM.
  - ✓ **One-vs.-all** approach.
  - ✓ Allows assigning segment to **multiple rooms**, or even **reject** segment.

SAD: Alternative Systems

- The 2-stage **proposed** system will be **evaluated** against **alternative** ones:

Baseline "room-independent" SAD system

- Build **2-class** (speech / non-speech) **GMM** for each room (1 mic. selected).
- Perform corresponding **Viterbi** decodings (one per room).
- Obtain **union** of resulting speech segments across rooms.

Contrastive 1: "Room-dependent" SAD with 3-state GMMs

- Build **3-class** ("inside" sp./"outside" sp./non-sp.) **GMM** for each room (1 mic.)
- Perform corresponding **Viterbi** decodings (one per room).
- **Purge** "outside speech" states to yield "room-dependent" SAD segments.

Contrastive 2: Two-step "room-dependent" SAD with MLPs [8]

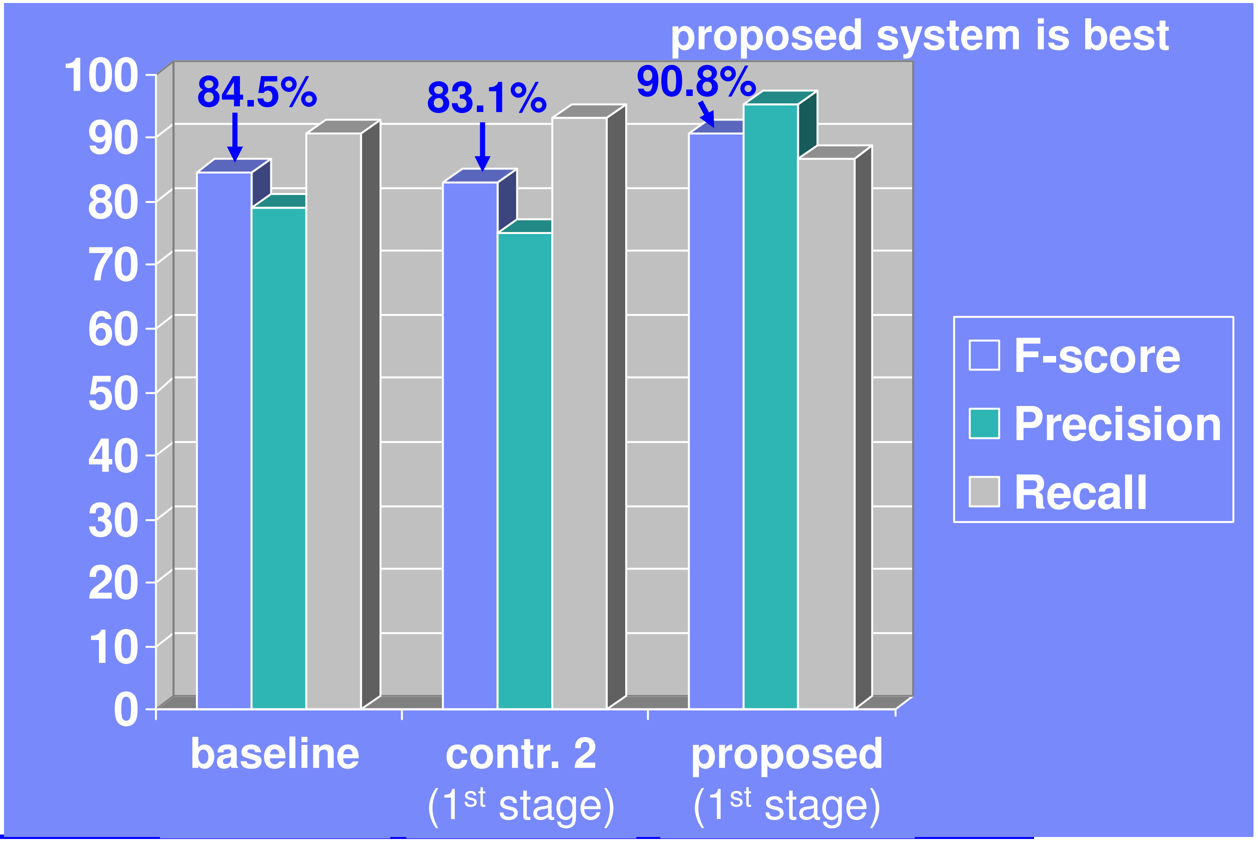
- Uses **MLPs** instead of GMMs.
- **1st step**: FSM decoder for each room mic., **majority voting** combination.
- **2nd step**: **EV**-based filtering of "outside speech" per room.

[8] Abad *et al.*, "The L²F system for the EVALITA-2014 speech activity detection challenge in domestic environments," *CLiC-it/EVALITA'14*.

SAD Results (I)

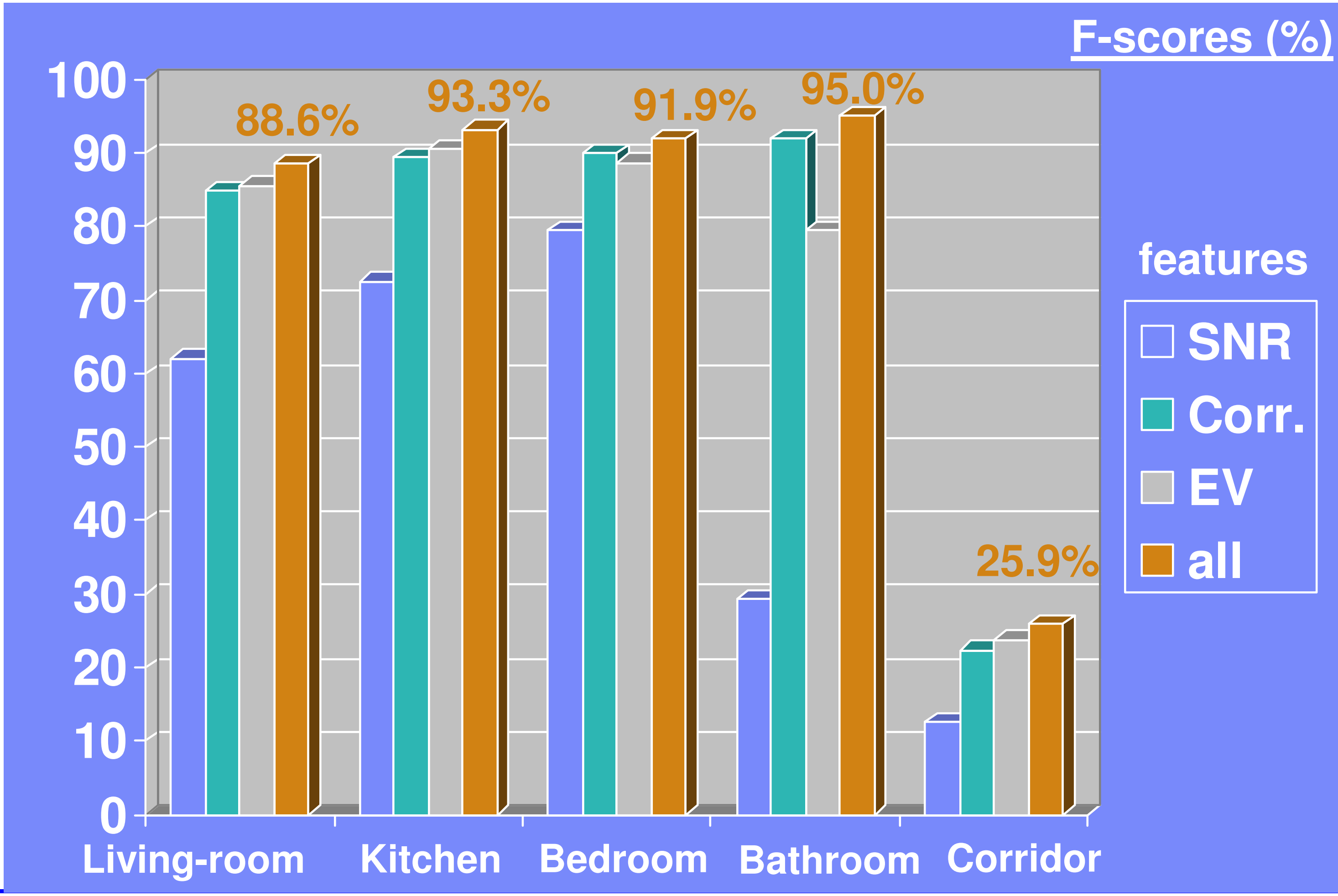
- **Experimental framework**:
  - ✓ Models (GMMs, SVMs) **trained** on "dev" set, **tested** on "test1"+"test2".
  - ✓ **Metrics: Frame**-based (10 ms) **precision**, **recall**, **F-score** (%).
  - ✓ **Results: DIRHA-sim** corpus.

1. **Evaluate "room-independent" SAD first:**



SAD Results (II)

2. **Evaluate room-selection feats. for "room-dependent" SAD:**
   - ✓ **Single** features (*R*-dim) vs. **all** features (3*R*-dim), for all ***R* = 5** rooms.
   - ✓ Feature fusion ("**all**") is best (features convey **complementary** info.).
   - ✓ **Corridor** performance is **worst** (located in the middle of apartment).



SAD Results (III)

3. Evaluate systems in "room-dependent" mode:

- ✓ As expected, "room-independent" SAD systems fail (low precision).
- ✓ Among "room-dependent" SAD systems, **proposed is best**.

System	F-score	Prec.	Recall
Contrastive 2 (MLP), 1st step only	40.92	26.29	92.31
Proposed (MS-GMM), 1st step only	49.27	35.32	81.47
Contrastive 1 (3s-GMM)	60.23	52.69	70.30
Contrastive 2 (MLP), both steps	57.61	48.22	71.56
Proposed (MS-GMM), both steps	$\boxed{74.46}$	68.50	81.58

Part II: BabyRobot Project

- **Motivation.**
- **Contributions – main ideas.**
- **Sensory setup.**
- **Perception system developed.**
- **DSR approach for Greek C&C.**
- **HRI evaluation scenario.**
- **Data.**
- **Results.**

-- **BabyRobot**: *Child-Robot Communication and Collaboration* (**babyrobot.eu**)

-- Tsiami *et al.,* “Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults”, *ICASSP’18*.

-- Tsiami *et al.,* “Multi3: Multi-sensory perception system for multi-modal child-robot interaction with multiple robots”, *ICRA’18*.

Motivation

- Increasing **popularity** of human-robot interaction (**HRI**) systems.
  - ✓ Driven by **advances** in robotic platforms and interaction technologies.
  - ✓ Wide range of **applications**, e.g. edutainment, assisted living, etc.
  - ✓ Multiple active **research projects**, e.g., **BabyRobot**, DE-ENIGMA, *etc*.
- Holy grail: **natural HRI, resemblance** to **human-human** communication.
  - ✓ Exchange of audio-visual information, crucially via **speech** & **gestures**.
  - ✓ HRI **perception**: speech & gesture recognition, localization (attention).
- Robot perception needs to be **robust** to:
  - ✓ **Noise** and **reverberation**.
  - ✓ Visual **occlusions** and **pose** variation.
  - ✓ **Complexity** of the audio-visual scene.
  - ✓ **Untethered**, **far-field**, **multi-party** interaction scenarios.
- **Challenging** to achieve by **robot-based sensing** alone.

Contributions (I)

1. **Pursue robustness** using **robot-external sensing**:
   - ✓ Multiple **audio-visual sensors** in the **far-field**.
   - ✓ Creates a "**smart-space**" for unobtrusive observation of the HRI scene.
   - ✓ Allows **fusion** of multiple audio-visual streams (inter- / intra-modal).
   - ✓ Perception becomes **robot-independent**.

➢ **Developed setup** employs **four Kinects (V2)**.

. . .

2. **Develop** three **perception modules** under this sensory setup, for:

➢ Multi-sensory audio-visual **speaker localization**.

➢ Multi-microphone **distant speech recognition**.

➢ Multi-view **gesture recognition**.

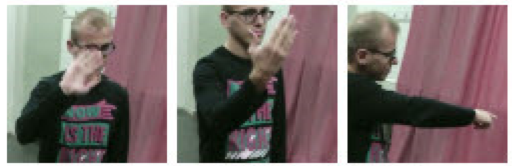
… adopting / integrating **standard techniques** from the literature.

Contributions (II)

3. **Modules** are developed for two user groups (**children** & **adults**):
   - ✓ Much interest on **cHRI**, but most components developed for adults.
   - ✓ User groups differ in interaction **behavior** & **articulatory** characteristics
- ➢ **Adaptation** and **training** schemes for the two user groups investigated.

Speech by a child

Gestures performed by a child



Gestures performed by an adult

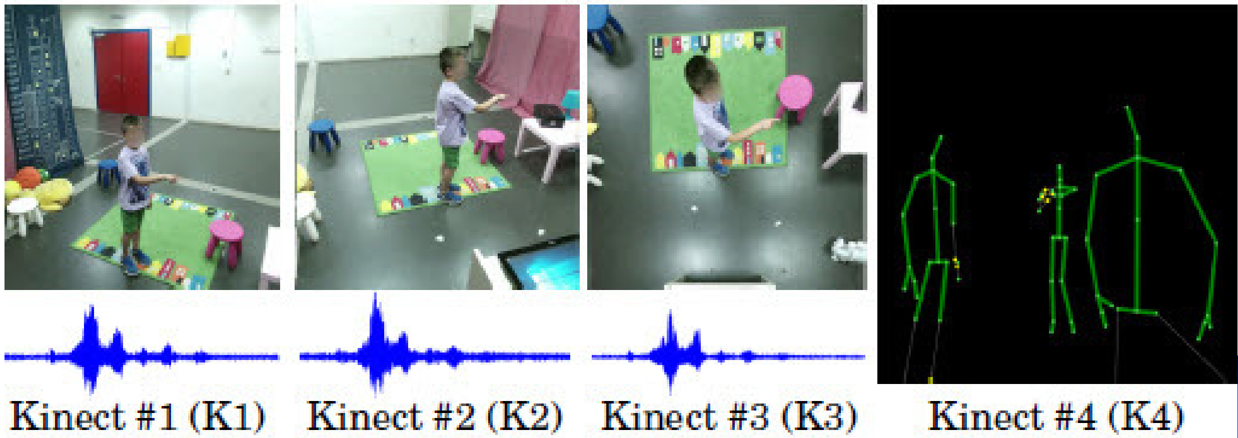
4. **Module integration** and **evaluation** within **use-case scenario**.

- ➢ **Integration** of perception modules within the **IrisTK** architecture.
- ➢ Development of a **“guess-the-object” HRI game** with a “**Furhat**” robot.
- ➢ Stand-alone **evaluation** of modules on children and adult data.
- ➢ **Evaluation of the HRI game** incorporating the integrated modules.

Sensory Setup (I)

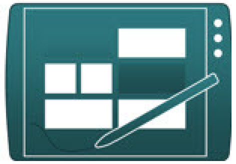
Four Kinects (V2 / Xbox One) are employed.

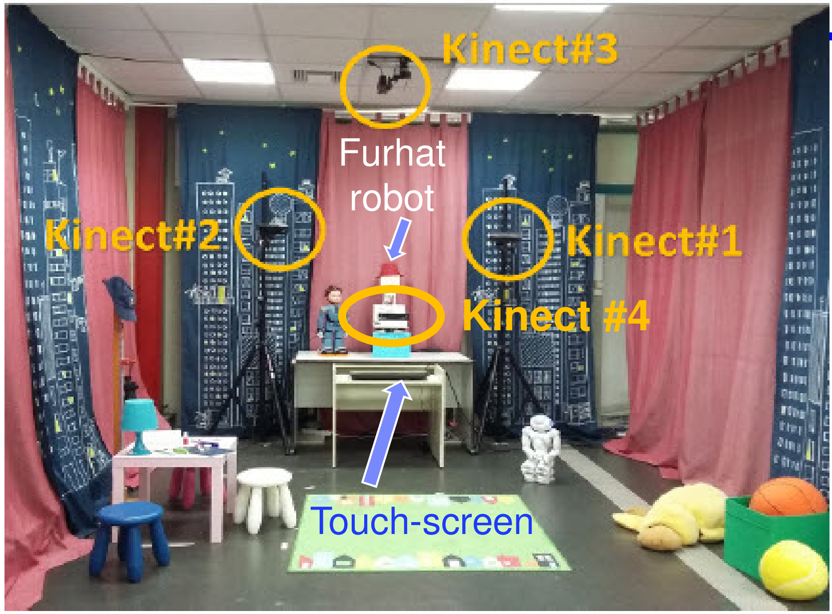
- **Three Kinects**, controlled by PCs running **Linux** (one master), provide:
  - ✓ **RGB video** (1920 x 1080 @ 30 fps).
  - ✓ **4 channels of audio** (16 kHz).
- **One Kinect** (controlled by a PC running **Windows**) provides:
  - ✓ Visual **skeleton** information (2D/3D coordinates of 25 joints @ 30 fps).
- **Unused** data streams:
  - ✓ RGB and audio channels of the fourth Kinect.
  - ✓ Depth streams of all.
- **Data streams** example (beamformed audio shown):



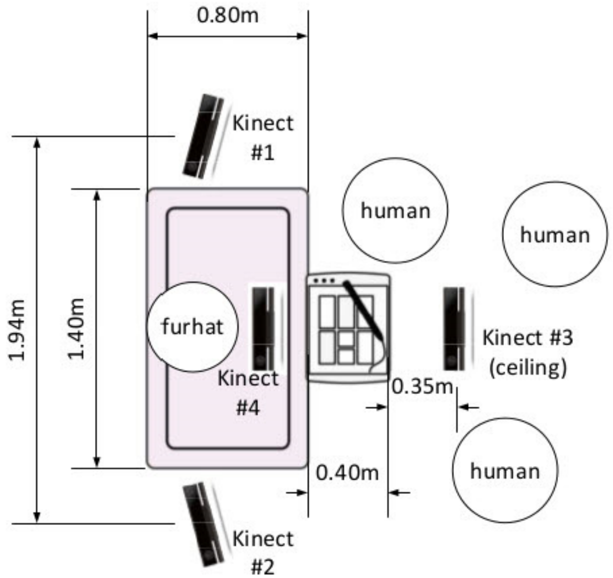
Sensory Setup (II)

- Sensors placed **indoors**, in a lab specially designed as a room for **cHRI**.
- Setup also involves:
  - ✓ "**Furhat**" robot.
  - ✓ **Touch-screen**.

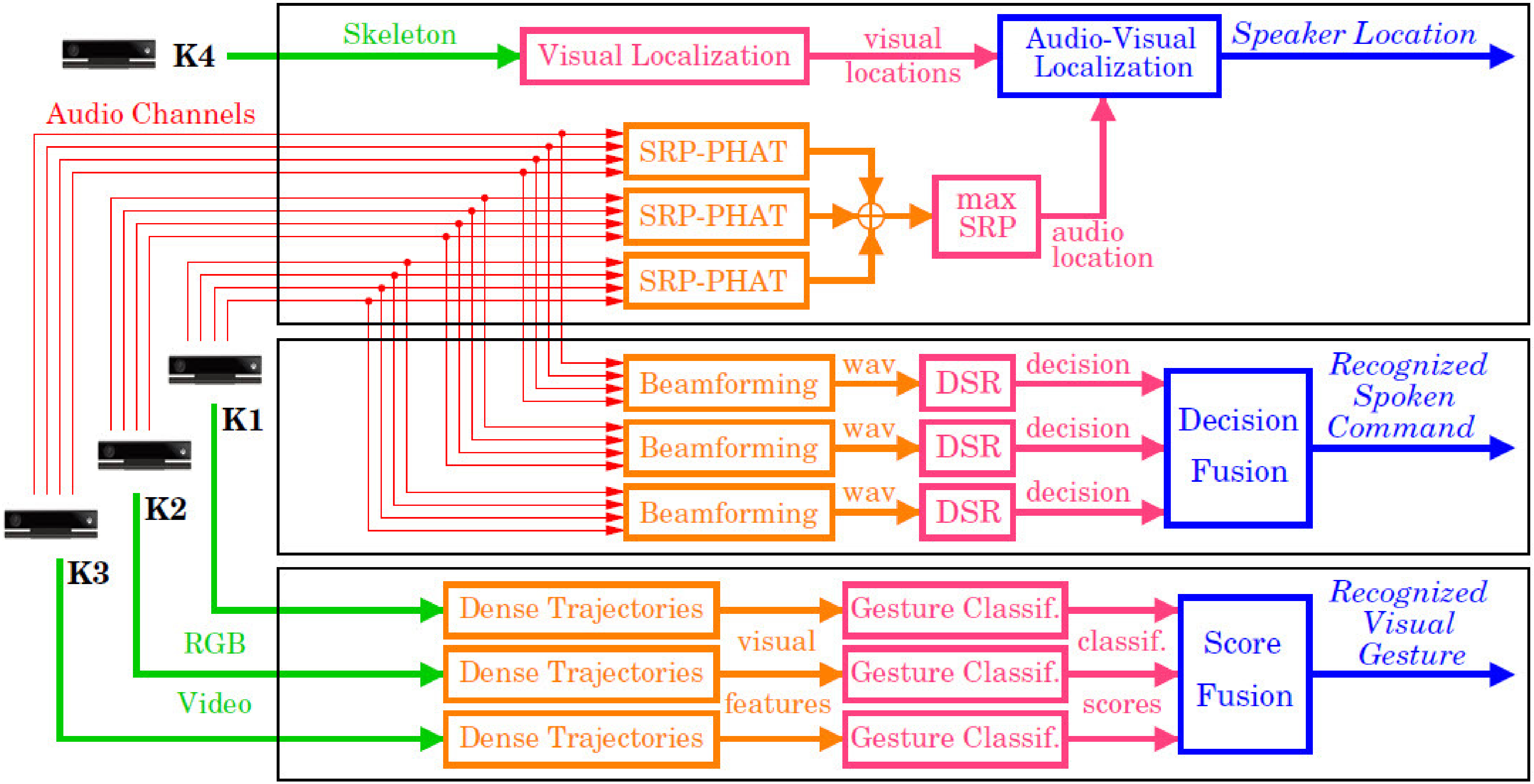




- Humans interact with robot in **confined HRI space** (scenario discussed later).
- **Kinects surround HRI area**:
  - ✓ K4 facing subjects.
  - ✓ K1, K2 at the sides.
  - ✓ K3 at the ceiling.
- Approximate **floorplan:**



Audio-Visual Perception System
Overview of 3 Modules



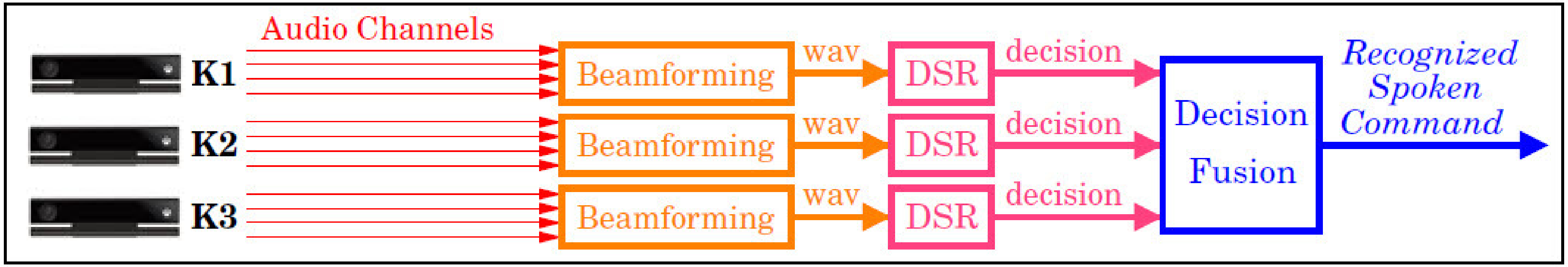
Signal processing blocks

Single-modal or -sensor decision blocks

Intra- or inter-modal fusion blocks

Distant Speech Recognition

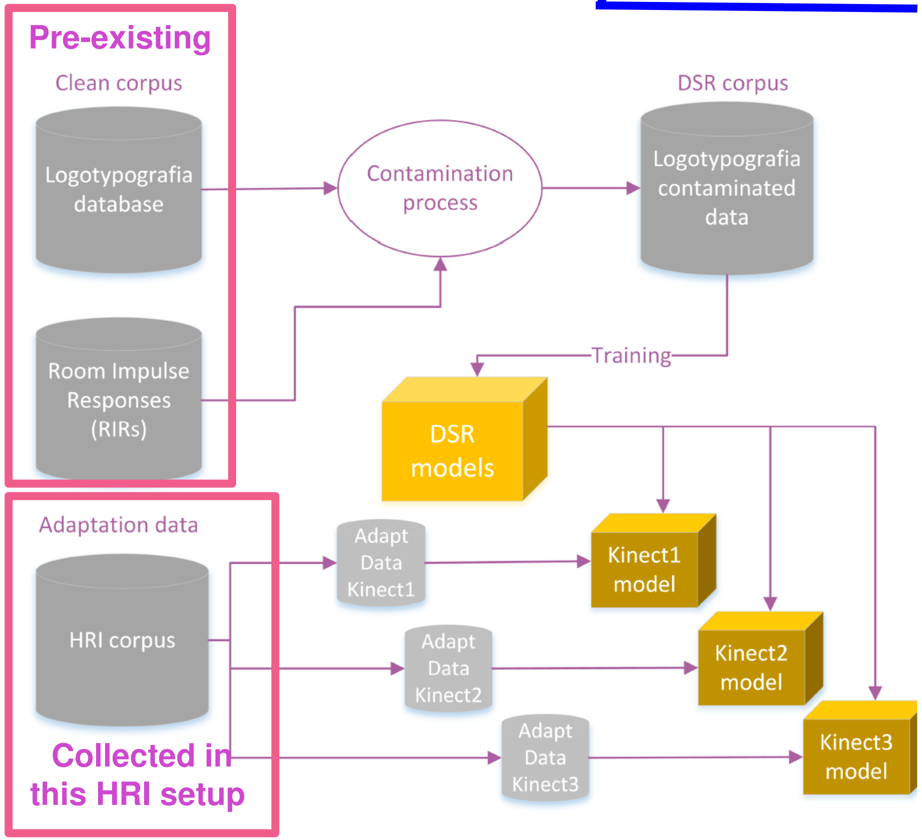
- Module **block diagram**. Utilizes **3 x 4 audio** Kinect channels.



- DSR system is **GMM-HMM** based, built on **HTK** for **Greek**. Main modules:
- **Beamforming** for intra-sensor signal fusion:
  - ✓ Simple **delay-and-sum** (no post-filtering).
- DSR **model training**:
  - ✓ **Contamination** of large available close-talking corpus.
  - ✓ Per Kinect MLLR **adaptation** based on HRI collected data.
- DSR **decoding**: **Grammar**-based due to simple HRI scenario (see later).
- Inter-sensor **decision fusion**: **Majority voting** of sensor results.

DSR Model Training

- "**Logotypographia**": Large, available Greek set (125 spk, 72 hrs, 50k wds).
- Close-talking part of it used (**22.6 hrs**)
- Contaminated with **RIRs** (T_{60} = 0.7 s), available from the DIRHA project, plus **white Gaussian noise**, simulating **far-field** conditions.
- **GMM-HMM** DSR system is trained:
  - ✓ Standard MFCC+derivs. frontend.
  - ✓ 3-state cross-word triphones(~8k)
  - ✓ 16 Gaussians per state.



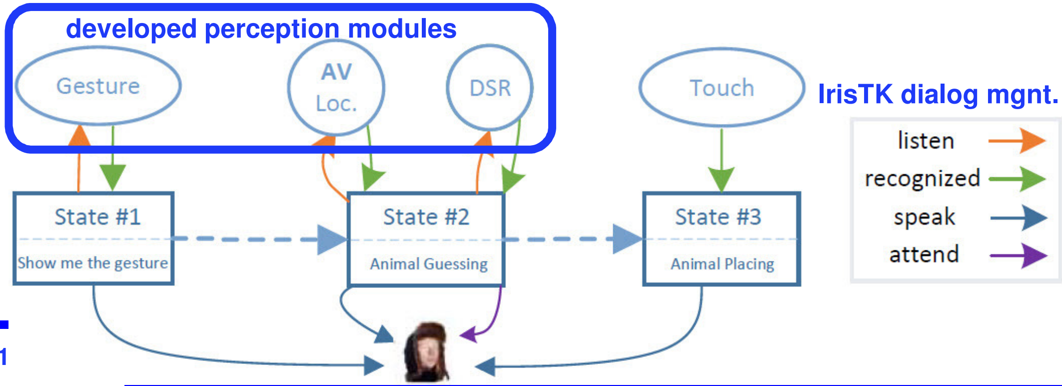
- Model **adaptation** follows, on data collected in the HRI setup:
  - ✓ For each Kinect sensor (#1, #2, #3).
  - ✓ Via **MLLR** (maximum likelihood linear transform).
  - ✓ Yields **3** adapted DSR models.

DSR Decoding and Fusion

- Viterbi decoding is **grammar-based**:
  - ✓ Helps in system robustness.
  - ✓ No understanding module needed.
  - ✓ Facilitated by simple HRI scenario adopted (see later).
  - ✓ Greek grammar consists of **~300 sentences**.
- System is "**always listening**":
  - ✓ DSR on running windows of **2.5 s** in duration, shifted by **0.6 s**.
  - ✓ After prompted by the dialog manager; "timed-out" after **5 s**.
- **Fusion** of recognition results:
  - ✓ Each Kinect array (3 in total) outputs a speech recognition hypothesis.
  - ✓ Fusion via **majority voting**.
  - ✓ In case of a **tie** (3 different results), user is prompted to repeat.

Use Case Scenario / HRI Game

- **Edutainment scenario**:
  - ✓ "**Guess-the-object**", within a "**form-a-farm**" HRI game.
  - ✓ Multiple **humans** (typically two) and a **robot** interact.
  - ✓ **Roles**: "**picker**" (picks the animal) and "**guesser**" (tries to guess it).
  - ✓ **19** animals, **5** characteristics (*e.g.*, color, size, number of legs, *etc.*).
- **HRI** unfolds in multiple "**states**" as follows:
  - ✓ **State 1**: "**Show-me-the-gesture**" determines roles.
    - If robot recognizes human gesture, it's the "picker", else "guesser".
  - ✓ **State 2** – Iterations (up to 5) of:
    - Guesser(s) trying to **identify** picked animal.
    - Picker providing **cues** (characteristic animal properties).
  - ✓ **State 3**: Human(s) **place** animal within farm drawn on **touchscreen**.



animal farm drawn on touchscreen

Data and Evaluation

- **Data collection**:
  - ✓ **Standalone data** for development and evaluation of perception modules
    - **20** adults; **28** children (ages **6-10**) [~ **1/3** female, **2/3** male].
    - In total: **3.7*k*** utts. (**~3 hrs**); **~400** gestures; **~1.6*k*** AV loc.“scenes”.
  - ✓ **Integrated HRI game data** for the evaluation of the entire system.
    - **12** pairs of adults.
    - **14** pairs of children.
    - **4 – 6** games for each pair.

- **Evaluation**:
  - ✓ Objective evaluation of **standalone perception modules (DSR)**.
  - ✓ Evaluation of **entire system** (HRI game).

DSR Evaluation (I)

- Evaluation focus lies on:
  - ✓ **Children** vs. **adult** performance.
  - ✓ Training and adaptation **strategies** for the two user groups.
- Strategies explored:
  - ✓ **No adapt** (speech only): Far-field Greek models by data contamination.
  - ✓ **Adults:** adaptation / training on adult data.
  - ✓ **Children:** adaptation / training on children data.
  - ✓ **Mixed:** adaptation / training on union of adult and children data.
- **4-fold cross-validation**.

DSR Evaluation (II)

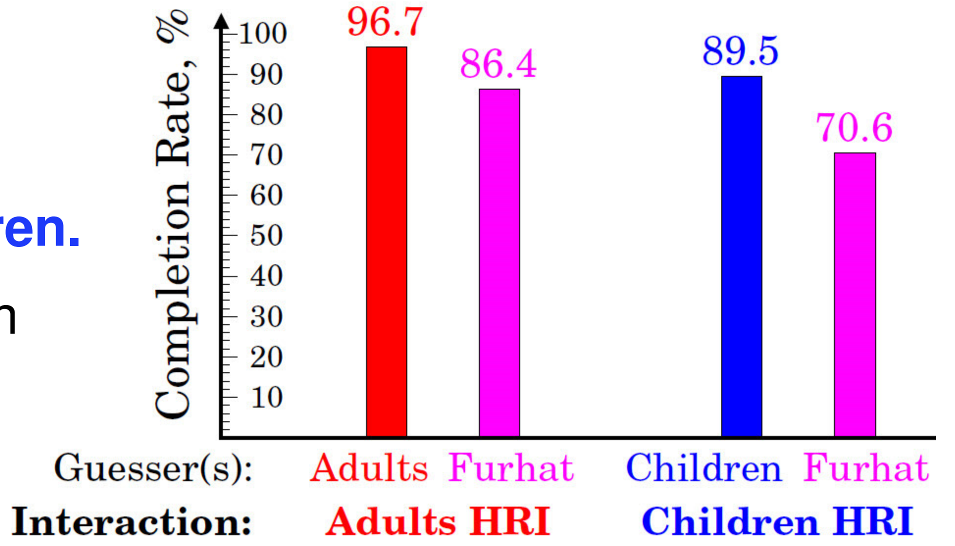
Test		DSR-Adaptation scheme: No-adapt	Adults	Children	Mixed
		Sentence Correct (%)			
Adults	K1	91.76	98.95	94.52	98.69
	K2	90.60	98.70	90.99	97.85
	K3	91.39	98.95	94.11	98.75
	Avg	91.25	**98.87**	93.20	98.43
	Fuse	92.41	**99.82**	94.42	99.77
Children	K1	70.53	72.31	95.95	82.95
	K2	72.48	73.85	95.95	82.52
	K3	66.83	67.63	94.60	80.70
	Avg	69.95	71.20	**95.50**	82.06
	Fuse	64.17	66.02	**98.97**	95.51

- Final recognition results are very satisfactory.
- User-group adapted/trained models perform well within group, poorly across.
- Mixed-group models are (near-)optimal for adults.
- Within-group modeling helps mostly for children.
- Fusion across Kinects helps.

Evaluation of HRI Game

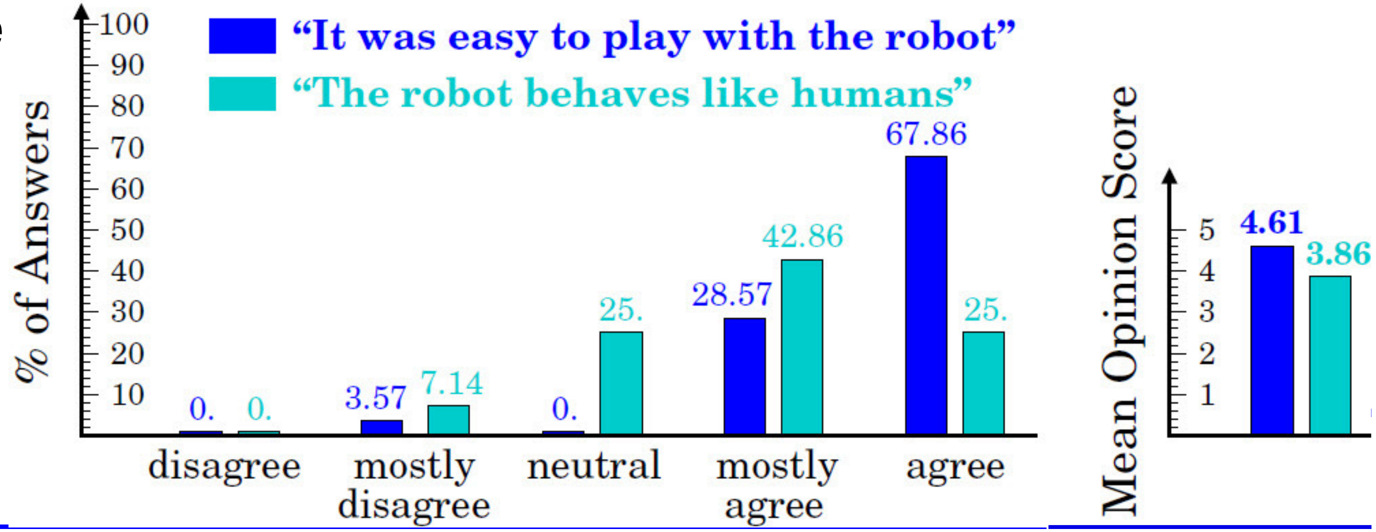
Online evaluation statistics:

- "**Guess-the-object**" successfully completed at **high rates**.
- **Adults** **better guessers** than **children.**
- **Furhat** is more "**fair**" as "picker" than humans (**adults** & **children**).



Subjective evaluation:

- **Children** rated the HRI **highly**.
- **Caveat: Children** "**ceiling effect**".



Conclusions

- Developed **command-based DSR** in **Greek** in **multi-microphone** smart environments for:
  - ✓ **Multi-room smart-home control** (**DIRHA** project).
  - ✓ **Child-robot interaction for edutainment** (**BabyRobot** project).
- Algorithmic details presented for various system **modules**.
- Evaluation on **real** and **simulated** data.
- Focus on **children** and **adult** user groups.
- **Satisfactory DSR results** obtained, demonstrating the importance of **fusing** multiple microphones in the **far-field**.

Acknowledgments

- Niki **Efthymiou**
- Panagiotis **Filntisis**
- Panagiotis **Giannoulis**
- Athanasios **Katsamanis**
- Petros **Koutras**
- Isidoros **Rodomagoulakis**
- Antigoni **Tsiami**

- **DIRHA** EU project consortium.
- **BabyRobot** EU project consortium.

Support by EU Horizon 2020 Project **BabyRobot**, under grant agreement no. 687831.