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

Gerasimos Potamianos¹,³ Petros Maragos²,³

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:

- 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)


**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).
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.

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.
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).

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

Results

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

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIRHA-sim</td>
<td>29.3</td>
<td>38.7</td>
</tr>
<tr>
<td>DIRHA-real</td>
<td>45.0</td>
<td>60.0</td>
</tr>
<tr>
<td>ATHENA-real</td>
<td>59.7</td>
<td>76.6</td>
</tr>
</tbody>
</table>

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

---

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 $r$: $x^r_m(t), \; m = 1, \ldots, M_r, \; r = 1, \ldots, R$
  - All signals at time $t$: $x_t = [x^1_1(t), \ldots, x^1_{M_1}(t), \ldots, x^R_1(t), \ldots, x^R_{M_R}(t)]$
  - Observation sequence of duration $T$: $X = [x_1, \ldots, x_T]$

- **“Room-independent” SAD:**
  - Find state seq. $Q' = [q_1, \ldots, q_T]$ that maximizes prob. $p(Q' | X)$

- **“Room-dependent” SAD:**
  - Find seqs. $Q^r = [q^1_1, \ldots, q^1_T], \ldots, [q^R_1, \ldots, q^R_T]$, for each room $r = 1, \ldots, R$,
  - that maximize $p(Q^1, \ldots, Q^R | X)$
Proposed “room-dependent” SAD system consists of two stages.

1\textsuperscript{st} stage: “Room-independent” SAD
- GMMs trained for each microphone.
- Fused by multi-stream framework.
- Viterbi decoding provides candidate speech segments to 2\textsuperscript{nd} stage.

2\textsuperscript{nd} stage: Inside/outside classification
- 1\textsuperscript{st}-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: $\lambda^r_m, m = 1, \ldots, M_r, r = 1, \ldots, R$.

- Fuse all GMM log-likelihoods in multi-stream style [7]:

$$L(q_t | x_t) = \sum_{r=1}^{R} \sum_{m=1}^{M_r} \log p(q_t | x^r_m(t); \lambda^r_m)$$

- 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).

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

SAD – 2\textsuperscript{nd} Stage: Overview

- **Main idea:**
  - For each “room-independent” speech segment (from 1\textsuperscript{st} stage);
  - For each room;
    - Compute “room-selection features” of the segment;
    - To discriminate if it originates from inside vs. outside room.

- Need discriminating features. Note that “outside” vs. “inside” speech has:
  - Lower energy \(\Rightarrow\) use SNR-based measurements.
  - Higher reverberation \(\Rightarrow\) use signal correlation, envelope variance.

- We employ three features – their histograms show room-discriminability:
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:

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

- then, concatenate them over all rooms:

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

Classification as “room-inside” or “room-outside” segment:

- Use room-specific, 2-class SVMs, trained on 3R-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.

**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:

![Bar chart showing performance metrics for different systems](chart.png)

- Baseline: 84.5%
- Contr. 2 (1st stage): 83.1%
- Proposed system: 90.8%

**proposed system is best**
## SAD Results (II)

### 2. Evaluate room-selection feats. for “room-dependent” SAD:

- **Single** features ($R$-dim) vs. **all** features ($3R$-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).

![Graph showing F-scores for different rooms and features](attachment:image.png)
### 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**.

<table>
<thead>
<tr>
<th>System</th>
<th>F-score</th>
<th>Prec.</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contrastive 2 (MLP), 1st step only</td>
<td>40.92</td>
<td>26.29</td>
<td>92.31</td>
</tr>
<tr>
<td>Proposed (MS-GMM), 1st step only</td>
<td>49.27</td>
<td>35.32</td>
<td>81.47</td>
</tr>
<tr>
<td>Contrastive 1 (3s-GMM)</td>
<td>60.23</td>
<td>52.69</td>
<td>70.30</td>
</tr>
<tr>
<td>Contrastive 2 (MLP), both steps</td>
<td>57.61</td>
<td>48.22</td>
<td>71.56</td>
</tr>
<tr>
<td>Proposed (MS-GMM), both steps</td>
<td>74.46</td>
<td>68.50</td>
<td>81.58</td>
</tr>
</tbody>
</table>
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](https://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.

![Images of child and adult gestures and speech]

Speech by a child          Gestures performed by a child          Gestures performed by an adult

greeting “come closer” pointing

greeting “come closer” pointing

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.

  ![Block Diagram](image)

- **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.
"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.
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.
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: 3.7k utts. (~3 hrs); ~400 gestures; ~1.6k 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.
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.