SLOC_SAD Evaluation tool 1.0: user guide

version 1.1 19/12/2013
For any doubt or request please contact Alessio Brutti (brutti@fbk.eu).

1 Introduction

This document describes the evaluation tool for the localization and speech activity detection task of the DIRHA special session at HSCMA 2014. Given the multi-room domestic scenario addressed in the DIRHA project, the goal of the task is to identify time boundaries, room and spatial coordinates of each speech event. A detailed description of the data and of the task is available at http://dirha.fbk.eu/hscma.
The scoring tool jointly evaluates Source LOCalization (SLOC) and Speech Activity Detection (SAD) performance. Scores are derived from those adopted under the CHIL project for the CLEAR evaluations[1].
The scoring tool is a C++ executable compiled under Linux.

1.1 Scenario Overview

The general scenario addressed in the DIRHA project refers to an apartment monitored by 40 microphones, distributed on the walls and the ceiling of its five rooms as shown in Figure 1. The consortium has created a set of real data as well as simulated data based on real room impulse responses, to evaluate a variety of signal processing algorithms:

- In the case of simulated data, several scenes, whose duration is 1 minute, were generated using a probabilistic framework. Each scene consists of a set of utterances and of other acoustic events produced in different rooms and positions. A variety of background noises, typical of the domestic scenario, are overlapped (i.e. added) to the scene.
- Conversely, real data are excerpts of a Wizard-of-Oz data collection. Each real scene includes a moving speaker, a rather quiet background, and system audio messages played through a ceiling loudspeaker. The speakers were located only in the Kitchen and Livingroom.

[1] http://clear-evaluation.org/

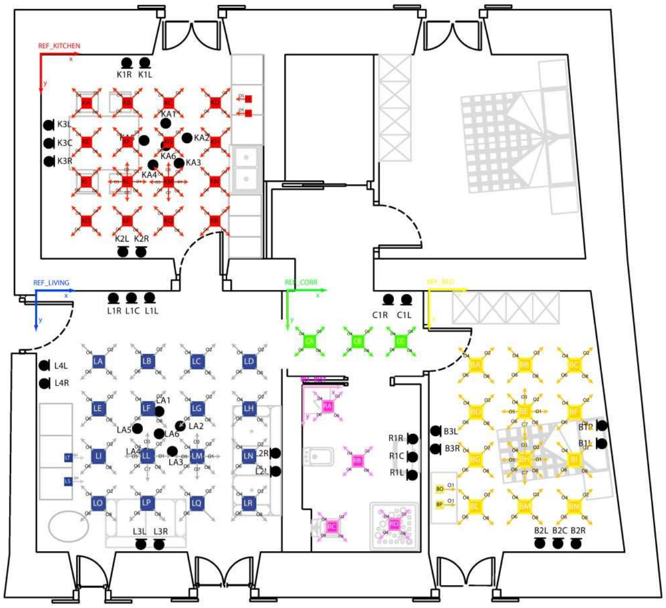

Fig. 1: Layout of the experimental set-up for simulated data. Squares and arrows indicate the possible positions and orientations of acoustic events.

For what concerns ground-truth information of the real data to use as reference, speech event segmentation was derived manually, while the speaker position coordinates were obtained using the infra-red tracking tool of multiple Kinect devices. For more details, please refer to the special session website http://dirha.fbk.eu/hscma, where short audio examples are also available.

1.2 Task Overview

The goal of the joint SLOC+SAD task is to detect, for each scene and for each room, the presence of speech events, and to locate the position of the sources producing such events. Speech events occurring in other rooms must be neglected, any other noise event must be neglected as well. The current evaluation considers only events generated in the Livingroom and Kitchen, although speech and noise events may occur anywhere in the apartment.

For each scene, a hypothesis file, compliant with the format described in Section 3, must be produced for each room. Further details are available at http://dirha.fbk.eu/hscma.

2 Reference Labels

Reference speaker positions and speech activities are reported every 50 ms in a reference file, together with the annotation of other acoustic events occurring in the 5 rooms. For **each scene** and **each room**, a reference file is provided consisting of a sequence of lines with the following format:

```
<time[s]> <n. sources in the room> <n. sources in other rooms>
<n. background noises> <event label> <x[mm]> <y[mm]> <z[mm]>
<labels of other events>
```

For each room, a specific coordinate reference system is defined (see Figure 1).

The example below refers to the kitchen. It represents a case in which a speech event (a phonetically rich sentence) is being produced in the room under analysis (second field set to 1), while the hair dryer is producing noise in the bathroom (third field set to 1). The coordinates of the speaker are (x=690,y=720,z=1500).

```
39.20 1 1 0 sp_ph_rich 690 720 1500 #+Hair_dryer(BATHROOM)
39.25 1 1 0 sp_ph_rich 690 720 1500 #+Hair_dryer(BATHROOM)
39.30 1 1 0 sp_ph_rich 690 720 1500 #+Hair_dryer(BATHROOM)
39.35 1 1 0 sp_ph_rich 690 720 1500 #+Hair_dryer(BATHROOM)
39.40 1 1 0 sp_ph_rich 690 720 1500 #+Hair_dryer(BATHROOM)
```

Conversely, in the second example below, still referring to the Kitchen, nobody is speaking in the targeted room (second field to 0), while the microwave is producing background noise (fourth field to 1). At the same time, a speaker is pronouncing a keyword in the Livingroom (third field to 1). Since the speech event is not occurring in the targeted room it must be discarded by the system.

```
10.55 0 1 1 - 0 0 0 #+bg_microwave_kitchen+sp_keyword(LIVINGROOM)
10.60 0 1 1 - 0 0 0 #+bg_microwave_kitchen+sp_keyword(LIVINGROOM)
10.65 0 1 1 - 0 0 0 #+bg_microwave_kitchen+sp_keyword(LIVINGROOM)
10.70 0 1 1 - 0 0 0 #+bg_microwave_kitchen+sp_keyword(LIVINGROOM)
```

3 Hypothesis format

For **each scene** and **each room**, the SLOC and SAD hypotheses must be delivered jointly in **a plain-text file** in the form of a stream of asynchronous speaker coordinates with the following format:

```
<time[s]> <x[mm]> <y[mm]> <z[mm]>
```

A speech event is defined by a continuous sequence of lines (including the time instant and the three spatial coordinates) with a minimum time resolution of 50 ms. The hypothesis production rate can also be not constant. The first and last time stamps of a continuous sequence represent the time boundaries for SAD evaluation. Any time gap larger than 50 ms will determine the instantiation of a new speech event. An example is reported below:

```
25.49 3075.0 3645.0 1550.0
25.54 3075.0 3645.0 1550.0
25.59 3075.0 3645.0 1550.0
25.64 3075.0 3645.0 1550.0
25.69 3075.0 3645.0 1550.0
25.74 3215.0 3685.0 1550.0
50.81 1925.0 2310.0 1550.0
50.86 1925.0 2310.0 1550.0
50.91 1925.0 2310.0 1550.0
50.96 1925.0 2310.0 1550.0
51.01 1925.0 2310.0 1550.0
51.06 1925.0 2310.0 1550.0
51.11 1925.0 2310.0 1550.0
```

where two speech events have been detected by the system:

- from time 25.49 s to time 25.74 s at coordinates (3075,3645,1550);
- from time 50.81 s to time 51.11 s at spatial position (1925 2310 1550).

NOTE:
the hypothesis file must be created even if there are no detected events in the scene: in that case, the file would be empty.
NOTE 2:
References are available with 50 ms temporal resolution. In case the system under evaluation works at a higher temporal resolution, the evaluation tool will “downsample” the hypotheses averaging the location estimates (see Figure 2). For lower working rates, the final output must have anyway a hypothesis advance step not larger than 50 ms (e.g. obtained by duplicating the most recent coordinate information or interpolating the hypotheses).

4 Evaluation procedure and Metrics

In the evaluation step, the hypothesis sequence and the reference file are compared one each other. For each reference line, the closest (in time) input line is selected in the hypothesis sequence and one of the four events below is generated:

- Deletion: no hypothesis available for a given reference line;
- False Alarm: an hypothesis is produced when there is no speech activity in the targeted room;
- Fine error: the distance between the estimated source position and the reference is smaller than 50 cm; the estimated location is considered possibly inaccurate but correct.

- Gross error: the distance between the estimated source position and the reference is larger than 50 cm; the localization algorithm has produced a wrong localization in space

Figure 2 provides an example for a deletion, a false alarm, a localization with time resolution equal to the one used in the references, and a localization at higher resolution which is downsampled averaging the source position estimates.

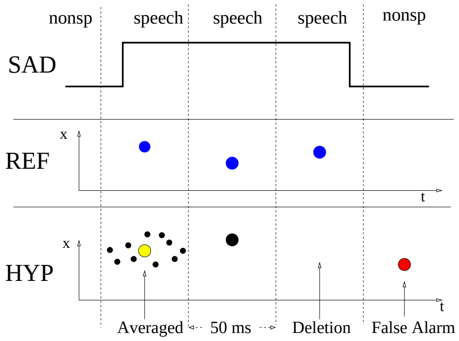

Fig. 2: Examples of behavior of the evaluation tool for the x coordinate for a single room: SAD is the boolean information of the speech activity, REF is the reference transcription of the x coordinate, HYP shows the results of the localization system in the case of output at higher frame rate, in the case of output at the evaluation rate and in cases of a deletion and a false alarm, respectively.

Given the classifications listed above, a series of metrics is computed to describe the behavior of the system under evaluation:

- Localization accuracy:
  - **Bias Error**: localization bias of fine errors.
  - **Bias Fine+Gross**: overall localization bias.
  - **RMSE Fine**
  - **RMSE Fine+Gross**
  - **Pcor**: number of fine errors over all the hypotheses produced. It basically measures the probability that a localization hypothesis is correct.
- Time boundaries accuracy:
  - **Deletion Rate**: number of missing hypotheses over all speech frames.
  - **False Alarm Rate**: number of false alarms over all non-speech frames.

- Detection performance:
  - **Precision** of the SAD component.
  - **Recall** of the SAD component.
  - **F score**.
- Overall detection error:
  - **Overall SAD Detection error**.
  - **Overall SAD+SLOC Detection error**.

In particular, the overall detection error (both for SAD and SAD+SLOC) aims at describing with a single number the performance of the system. It is computed as:

$$\text{SAD(+SLOC) Detection error} = \frac{N_{del} + N_{fa}(+N_{gross})}{N_{ref}},$$

where N_{del}, N_{fa} and N_{gross} are the total numbers of deletion, false alarms and gross errors respectively, while N_{ref} is the total number of references. Wherever possible, the results will be reported in a disaggregated fashion, differentiating among cases in which there are noises in the targeted room, interferers (noise or speech) in another room, background noises;

5 Scoring the hypothesis

The evaluation tool is a C++ executable. It is run with the following syntax: `./SLOC_SAD_Eval -list` *fileListName* `-totalSummary` *summaryFileName* It processes a list of hypothesis files and, for each of them (e.g each scene, each room), it produces two outputs:

- a file reporting the classification of each reference line;
- a summary file reporting the metrics described.

The list of files to process is provided in the plain-text file specified by the argument "-list". The file includes, on each line, the input file, the related reference file and the two output files. The example below considers two scenes (sim1 and sim2) both in the Kitchen and the Livingroom:

```
Hyp/sim1/Kitchen/output.hyp Ref/sim1/Kitchen.ref Output/sim1/Kitchen.out Output/sim1/Kitchen.sum
Hyp/sim1/Livingroom/output.hyp Ref/sim1/Livingroom.ref Output/sim1/Livingroom.out Output/sim1/Livingroom.sum
Hyp/sim2/Kitchen/output.hyp Ref/sim2/Kitchen.ref Output/sim2/Kitchen.out Outoput/sim2/Kitchen.sum
Hyp/sim2/Livingroom/output.hyp Ref/sim2/Livingroom.ref Output/sim2/Livingroom.out Output/sim2/Livingroom.sum
```

The tool computes also the average performance over all hypothesis files and delivers it in the file specified by the parameter "-totalSummary". An example of summary file is reported in table 1.

EVALUATION RESULTS	Overall	Noise in room	Noise outside	Background noise
Event type:	sp			
Bias fine (x,y,z)[mm]	(-14.8,-33.5,50.0)			
RMSE fine [mm]	232.7			
Bias fine+gross (x,y,z)[mm]	(-32.2,-99.7,58.5)			
RMSE fine+gross [mm]	1076.4			
Pcor	0.550[4378/ 7966]	0.336 [288/ 856]	0.612 [1316/ 2151]	0.509 [1439/ 2825]
Deletion rate	0.263 [2836/10802]	0.297 [362/ 1218]	0.285 [856/ 3007]	0.393 [1832/ 4657]
False Alarm rate	0.168 [14318/85278]	0.204 [2433/11900]	0.269 [8111/30107]	0.108 [4410/40981]
Loc. frames for error statistics	7966			
Overall SAD detection error	0.179			
Overall SAD+SLOC detection error	0.216			
Precision	0.461 [125/271]			
Recall	0.788 [149/189]			
Fscore(1.00)	0.582			
Total number of references	96080			

Tab. 1: Example of summary file in case of simulated data.

The parameters "-list" and "-totalSummary" are mandatory. For more parameters see the in-line help running the binary file without any argument.

2D vs 3D evaluation
The default evaluation considers the three spatial coordinates (x,y,z). However, estimating the height of the speaker may result quite hard, in particular given the adopted microphone deployment. The user can evaluate the system output in a 2D modality (neglecting the z-coordinate) by using the "-2D" flag when running the tool.
NOTE
Since in some scenes there are no events to be detected some of the single summary files may result empty.

5.1 Support scripts

In order to ease the use of the scoring software, the script `runEval.sh` is distributed. It supports the user in the creation of the evaluation lists and in running the evaluation tool.
To use it:

- produce the system outputs using the same file name for each room and scene and put the files in a folder structure parallel to that of the data;

- run `runEval.sh` *hypFolder hypName referenceFolder outputFolder summaryName*;
  where:

  - *hypFolder* is the name of the folder containing the hypothesis files;
  - *hypName* is the name of the hypothesis files;
  - *referenceFolder* is the name of the folder where the original data and the references are stored;

  - *outputFolder* is the name of the folder where the scores will be delivered;
  - *summaryName* is the name of the file containing the total summary.

- the script will create
  - in *outputFolder* a structure parallel to the input structure;
  - single evaluation files named *roomName.out* and *roomName.sum* (e.g. Kitchen.out, Kitchen.sum) and stored in the output structure;
  - a text file named *summaryName* with the overall scores.

For example, if we consider the simulated development data set:

- the development material is structured in this way:
  HSCMA_DIRHA_dev/Simulations/Dev/[GE|GK|IT|PT]/sim[1-10]/
  Signals/Mixed_Sources/[Livingroom|Kitchen];
- For each room and each session the hypotheses produced by the system to be evaluated are stored in a file name "output.hyp";
- The output generated by your system must be put in the following folder structure:
  Hyp/Simulations/[GE|GK|IT|PT]/
  sim[1-10]/[Livingroom|Kitchen]/output.hyp
- run the script as:
  ```
  sh HSCMA_DIRHA_dev/EvaluationTool/runEval.sh Hyp/Simulations
  output.hyp HSCMA_DIRHA_dev/Simulations Evaluation summary.txt
  ```
- the scores will be stored in the folder "Evaluation", while the total summary of the 80 files (i.e. 2 rooms x 40 sessions) will be saved in "summary.txt";