# The DIRHA Portuguese Corpus: A Comparison of Home Automation Command Detection and Recognition in Simulated and Real Data

**Miguel Matos**[1,2], **Alberto Abad**[1,2], **António Serralheiro**[1,3]
[1]L[2]F - Spoken Language Systems Lab, [2]IST - Instituto Superior Técnico, University of Lisbon, [3]Academia Militar
INESC-ID Lisbon, Lisbon - Portugal
jmatos, alberto, ajs@l2f.inesc-id.pt

**Abstract**

In this paper, we describe a new corpus –named DIRHA-L2F RealCorpus– composed of typical home automation speech interactions in European Portuguese that has been recorded by the INESC-ID's Spoken Language Systems Laboratory (L$^2$F) to support the activities of the Distant-speech Interaction for Robust Home Applications (DIRHA) EU-funded project. The corpus is a multi-microphone and multi-room database of real continuous audio sequences containing read phonetically rich sentences, read and spontaneous keyword activation sentences, and read and spontaneous home automation commands. The background noise conditions are controlled and randomly recreated with noises typically found in home environments. Experimental validation on this corpus is reported in comparison with the results obtained on a simulated corpus using a fully automated speech processing pipeline for two fundamental automatic speech recognition tasks of typical "always-listening" home-automation scenarios: system activation and voice command recognition. Attending to results on both corpora, the presence of overlapping voice-like noise is shown as the main problem: simulated sequences contain concurrent speakers that result in general in a more challenging corpus, while real sequences performance drops drastically when TV or radio is on.

**Keywords:** robust speech processing, real vs. simulated data, voice activity detection, speech recognition, keyword spotting, home automation

## 1. Introduction

In domestic environments, the presence of reverberation, background noise and interfering sources, critically degrades the performance of standard speech processing algorithms. A possible solution for improving the overall recognition robustness is the adoption of a network of distributed microphones that can partially reduce the impact of these nuisance factors on spoken dialogue home automation systems (Seltzer, 2003; Chu et al., 2006). This is the scenario addressed in the DIRHA project[1], in which an "always-listening" monitoring system is foreseen to be able to analyse and understand the activities and the intentions of the users inside a domestic environment. In the most typical DIRHA scenario, the automation system can be operated based on a dialogue manager strategy in which spoken interactions are usually initiated by the user by means of a certain activation sentence. Then, a spoken dialogue session is initiated, in which the user can provide specific home automation commands or ask for status information.

The availability of relevant corpora characterizing home environments is of fundamental importance to support research on the area of multi-channel processing for smart-home applications. Thus, some related data collection efforts have been conducted in the past. For instance, in the context of smart-room services for meeting and seminar assistance, in which typically a single room is simultaneously characterized (Janin et al., 2003; McCowan et al., 2005; Moreau et al., 2008). Concerning multiple-room environments, in (Vacher et al., 2014) automation commands in French were recorded in a 5-room smart home, overlapping either with noises or background events. Within the DIRHA project, some home automation control speech databases have been also collected including both activation sentences and commands for home automation. In (Tsiami et al., 2014), a Greek speech database with real multi-modal data in a two-room environment is described. The corpus described in (Cristoforetti et al., 2014) consists of a challenging database containing simulated data in four different languages: Italian, Austrian German, Greek and European Portuguese.

In this paper, we describe the data collection protocol conducted at the INESC-ID's Spoken Language Systems Laboratory (L$^2$F) to build the new multi-channel and multi-room corpus composed of typical home-automation interactions in European Portuguese, called DIRHA-L2F RealCorpus. Experimental validation on this corpus is reported in comparison with the results obtained on the simulated corpus described in (Cristoforetti et al., 2014) –named DIRHA-L2F SimCorpus– using a fully automated speech processing pipeline for two typical recognition tasks in multi-room multi-channel acoustically challenging environments: system activation and voice command recognition. Both the real and simulated corpora characterise different household multi-room environments equipped with a large number of microphones. The two main components of the processing system (the multi-room speech activation module and the multi-channel automatic speech recognition system) make extensive use of channel selection methods for fully exploiting the large number of microphone channels present in each corpus.

This paper is organised as follows. Section 2. describes the DIRHA-L2F Corpora under study, including both the new real and the simulated data sets. Section 3. introduces the speech processing pipeline used in this work for validation of home automation command detection and recognition. Experimental validation is presented and discussed in Section 4.. Finally, Section 5. concludes the paper with final remarks and future work.

[1]https://dirha.fbk.eu

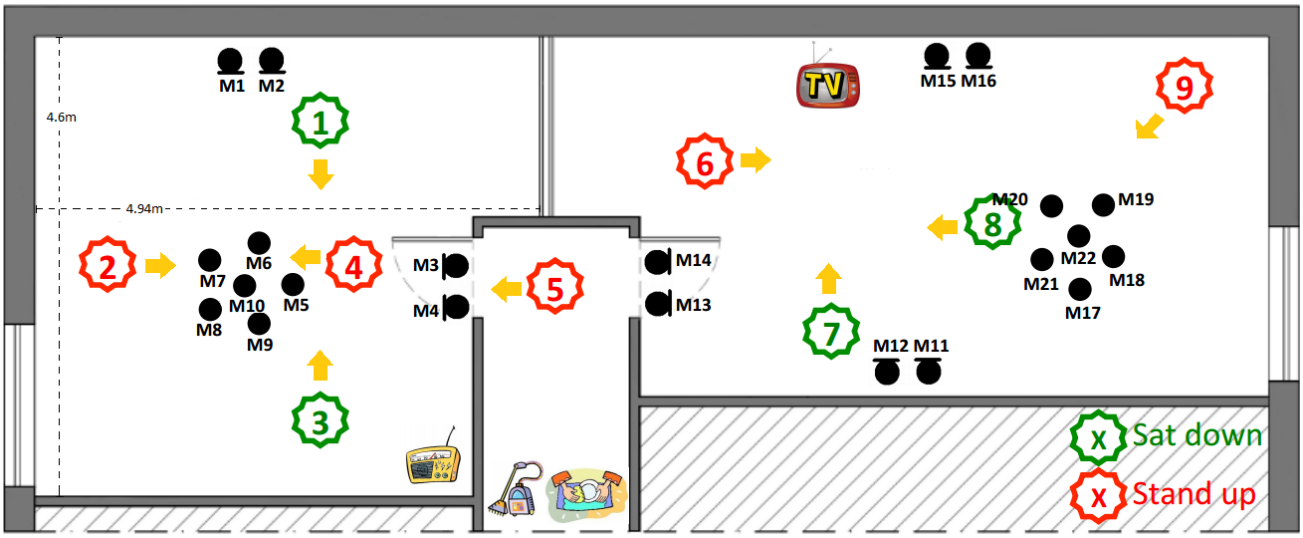


Figure 1: The DIRHA-L2F room set-up: microphones, speaker and noise interference positions and directions used in the DIRHA-L2F RealCorpus collection.

## 2. The DIRHA-L2F Corpora

### 2.1. The DIRHA-L2F RealCorpus

The DIRHA-L2F RealCorpus is a multi-microphone and multi-room database collected at the INESC-ID's Spoken Language Systems Laboratory ($L^2F$). Figure 1 shows the floorplan of the DIRHA-L2F recording environment composed of two office rooms and one corridor connecting them. The collected data contains real recordings in European Portuguese language of 20 gender balanced speakers, ageing between 20 and 60. For each speaker, 12 sessions of approximately 1-2 minutes were collected, totalling 240 sessions which corresponds to about 4.5 hours. During the first 9 sessions, speakers were placed at pre-defined constant positions and directions (see Figure 1), while during the last 3 sessions, speakers were allowed to move freely. Each session contains a read phonetically rich sentence, a read keyword sentence followed by a read command and a spontaneous command (sometimes accompanied of an spontaneous keyword activation sentence).

**Recording equipment** The DIRHA-L2F recording environment is equipped with 22 omni-directional Shure MX391-O microphone channels distributed among the two office rooms as shown in the Figure 1. Moreover, two Shure WH30TQG headset microphones with wireless connection are available to collect reference close-talking speech. All channels are recorded with a sampling rate of 48kHz and 16-bits quantization using the Focusrite Octopre MKII Dynamic microphone pre-amplifiers and A/D converters. The collected signals are digitally received by the computer through the RME HDSPe RayDAT board and saved in WAV-PCM format. All channels are synchronized using the clock signal generated by expansion word clock module board RME HDSP 9632.

**Recording protocol** During the recording sessions, speakers hold a tablet from which they received instructions including the position to occupy at each session or the next sentence to read/produce. The instructions were shown in the form of slides including text and images and the prompts were randomly generated for each speaker. The slide-show was remotely controlled by a recording monitor so that it was possible to control the amount of silence between events. In the particular case of read activation sentences and read commands, the text to read was shown next to an image of a room with an arrow aiming at the object corresponding to the action. Figure 2 shows an example corresponding to the Portuguese home control command *Acende a Luz / Turn on the light*. For spontaneous commands, the same kind of images with arrows were shown without including text.


Figure 2: Example of slide used for recording a read activation sentence followed by a read command. In English: *DIRHA start session ... turn on the light.*

**Noise environment conditions** The acoustic environment for each session was controlled and one of the following background noise conditions was randomly recreated: quiet, air conditioning, open window, vacuum cleaner, radio, TV, and kitchen noises. Table 1 shows the number of

times each condition is present in the complete corpus (in some sessions more than one noise condition is present simultaneously). The radio, TV, vacuum cleaner, and kitchen acoustic conditions were recreated using loudspeakers located in fixed positions marked in Figure 1. In the case of radio and TV, on-line available programs from local broadcasters were reproduced. In the case of kitchen noises and vacuum cleaner, real recordings made at home were played.

| Acoustic Condition | | | | | | |
|---|---|---|---|---|---|---|
| Quiet | AC | OW | TV | Radio | Kitchen | Vacuum |
| 60 | 56 | 43 | 40 | 20 | 20 | 20 |

Table 1: Number of recording sessions in the DIRHA-L2F RealCorpus per acoustic condition: Quiet, Air Conditioning (AC), Open Window (OW), TV, Radio, Kitchen and Vacuum cleaner (Vacuum).

**Annotation protocol** The close-talking speech channel was used to ease the transcription process. First, automatic speech/non-speech segmentation was applied to detect speech events on the head-set channel. Then, a human annotator manually corrected the generated segments, checked for possible mispronunciations in the read speech events, and transcribed the spontaneous speech events. Next, a different transcription was generated for each room. To this end, the speakers were instructed to hit two wooden spoons as close as possible to the microphone head-set to produce an impulsive sound at the beginning of each session. The delay of this impulsive sound at each microphone with respect to the head-set channel was semi-automatically computed and used to adjust the segment boundaries of the two far-field transcriptions per session. Notice that these resulting segmentation labels are only correct for the first 9 sessions in which the speaker is at a fixed position. For that reason, for each session a representative channel of each room was manually inspected to eventually correct segmentation labels. In this final manual transcription step, uncontrolled noises occurring during the session were also annotated. In addition to the transcriptions, the database includes information of the noise event files that were played at each session, which can be useful for instance to perform noise cancellation experiments. Moreover, room impulse responses corresponding to approximately the 9 fixed positions and orientations of the real database recordings were measured using the Exponential Sine Sweep Method (Farina, 2000; Ravanelli et al., 2012) and are included in the database.

### 2.2. The DIRHA-L2F SimCorpus

The DIRHA SimCorpus (Cristoforetti et al., 2014) is a multi-microphone and multi-language database containing simulated acoustic sequences derived from the microphone-equipped apartment located in Trento (Italy), named ITEA. The simulated corpora for the different languages –including European Portuguese (EP)– were produced thanks to a technique that reconstructs, in a realistic manner, multi-microphone front-end observations of typical scenes occurring in a domestic environment. For each language, the corpus contains a set of acoustic sequences of duration 60 seconds, at 48kHz sampling frequency and 16-bit accuracy, observed by 40 microphone channels distributed over five rooms, Figure 3.

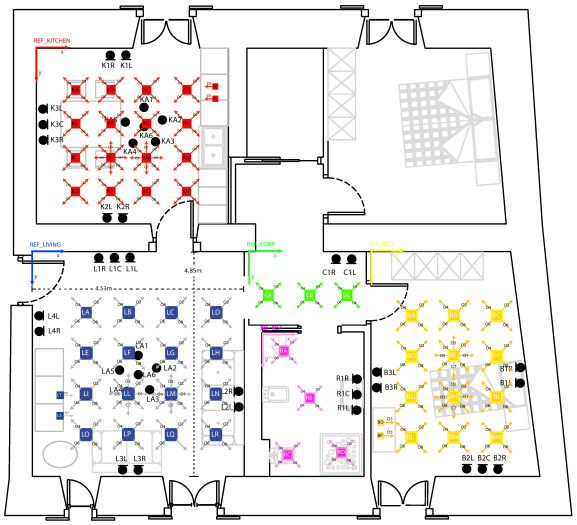

Figure 3: The ITEA apartment floor-plan with the speaker positions and orientations used for the simulation of the DIRHA SimCorpus.

For the DIRHA SimCorpus in EP, hereinafter referred to as DIRHA-L2F SimCorpus, a clean-speech data set of very high quality close-talking speech signals was collected to derive the simulated corpus. The data set contains 20 speakers with an equal gender distribution, ageing between 25 and 50. The EP simulated corpus is divided into two chunks (*dev* and *test*) containing 75 acoustic sequences each, with 10 different speakers in each data set (Cristoforetti et al., 2014).

## 3. Multi-channel speech processing pipeline for home applications in Portuguese

The DIRHA-L2F RealCorpus is analysed based on experimental tests conducted with a baseline processing system designed to perform the two detection and recognition tasks in domestic environments mentioned previously: system activation and voice command recognition. For each task the proposed processing pipeline is formed by two distinct modules that at the same time integrate several technological components: 1) multi-room speech activity detection (SAD), and 2) multi-microphone key-phrase spotting or command recognition. Figure 4 shows the main blocks that constitute the processing pipeline for testing the corpus.
Both modules make extensive use of channel selection strategies to exploit the available network of microphones (Obuchi, 2004; Jeub et al., 2011; Wölfel et al., 2006). The advantage of this kind of approaches is that they do not require prior information about the microphone set-up and the speaker positions, contrasting to the classical microphone array beamforming methods (Masgrau et al., 1999; Abad, 2007; Lecouteux et al., 2011).

### 3.1. Multi-room SAD

The multi-room SAD module takes as input the audio streams of all the microphones in the apartment and gen-

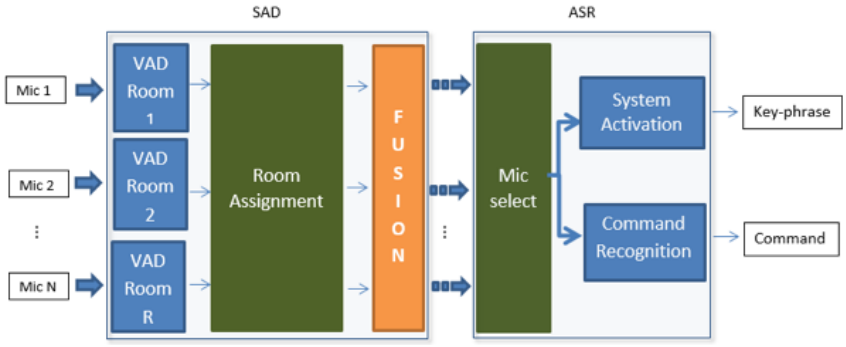


Figure 4: Block diagram of the L$^2$F multi-room and multi-channel speech processing pipeline for system activation detection and voice command recognition.

erates a single speech/non-speech segmentation for the entire home. It combines multi-channel model-based speech classification with automatic room localization.

First, single-channel speech/non-speech segmentation is run individually for each channel of the house with a classification block, which is implemented using an artificial neural network of the multi-layer perceptron (MLP) type (Meinedo, 2008), based on perceptual linear prediction (PLP) features. Then, the SAD module smooths the probabilities obtained by the classifiers, which are finally used by a finite state machine to obtain "speech" and "non-speech" segments. A baseline room-dependent segmentation is obtained based on the majority voting fusion of all the channel segmentations resulting from the microphones of that specific room. Speech segments detected at each room are time-aligned to obtain possible candidate speech events simultaneously detected in several rooms. Then, all candidate speech events are further processed to decide in which room they were originated, which we refer to as "room assignment" stage. Room localization information is obtained based on the envelope variance (EV) (Wolf and Nadeu, 2010) measures of each channel. More details about this module can be found in (Abad et al., 2014).

### 3.2. Multi-room ASR

The multi-room ASR module (Matos et al., 2014) takes as input the multiple-channel audio segments provided by the SAD module and selects the most convenient channel for further processing, which will consist either on key-phrase spotting or voice command recognition depending on the aimed task indicated by the dialogue manager. For both system activation and command recognition tasks the same set of acoustic models (AM) and multi-channel processing strategies have been considered.

**Multi-channel processing** The EV measure is used for microphone selection, since it presents a good trade-off between performance for ASR applications and processing complexity (Wolf and Nadeu, 2010; Wolf, 2013; Matos et al., 2014). For each candidate speech segment, the selected channel for automatic speech recognition is the one that obtains maximum EV among all the possible microphone channels.

**Acoustic models** A data-simulation based approach was adopted to create training data for the characteristics of the DIRHA far-field environment, by artificially convolving "clean" recordings from the BD-PUBLICO corpus (Neto et al., 1997) with IRs measured at a number of locations in the ITEA apartment, and contaminating them with noise at various SNR values. Then, training data representing different acoustic conditions were used together to obtain a single set of multi-condition acoustic models. The AM consists of word-internal tied-state context-dependent triphones of 3 states and 16 Gaussians per state. Feature characterization is based on the conventional 13-dimensional MFCCs, augmented by their first and second derivatives, and mean normalized, thus reaching a dimensionality of 39. More details about acoustic model training can be found in (Matos et al., 2014).

**ASR tasks** For the system activation task, a classical keyword-filler approach has been considered. The background HMM is trained using all the AM training corpora and it consists of a 24-states HMM with left-to-right transitions and 32 Gaussian mixtures per state. For the command recognition task, an equally-likely finite state grammar formed by all the unique possible command sentences was initially used. However, in practice it was necessary to use an extended command grammar incorporating the background model to better handle inaccurate segmentations provided by the automatic SAD. Moreover, a different grammar is considered for the simulated and real test data sets.

## 4. Experimental validation: Real vs Simulated data

The processing chain for system activation and voice command recognition of home automation applications in European Portuguese (EP) is used for validation and comparison of simulated and real multi-room and multi-channel in-domain data. The system activation task evaluation indistinctly considers the detection of both read and spontaneous system activations (the latter are only present in the real data), while the voice command recognition evaluation is restricted to the detection and recognition of read spoken commands. In these experiments, the development set (*dev*) of the DIRHA-L2F SimCorpus has been used for parameter tuning of the systems, while results are provided for the *test* set of the DIRHA-L2F SimCorpus and for the DIRHA-L2F RealCorpus (there is only one partition of the real corpus).

To facilitate the analysis of results in the the DIRHA-L2F RealCorpus, recording sessions have been grouped into 4 categories: quiet background QB , low and stationary noise background LB (air conditioning and open windows), high and non-stationary noise background HB (vacuum cleaner and kitchen noises) and voice-like noise background VB (radio and TV).

### 4.1. Key-phrase Spotting Evaluation

Tables 2 and 3 present F-score (%) results of detected system activations using ground-truth and automatic segmentation, respectively. In both cases, the multi-channel key-phrase spotting component processes all the segments either labelled (ground-truth) or classified (automatic segmentation) as speech.

| Corpus | Set | Prec. (%) | Recall (%) | F-score (%) |
|---|---|---|---|---|
| SimCorpus | test | 76.19 | 64.00 | **69.57** |
| RealCorpus | QB | 92.22 | 81.37 | 86.46 |
| | LB | 93.10 | 78.83 | 85.38 |
| | HB | 97.78 | 68.75 | 80.73 |
| | VB | 94.03 | 61.76 | 74.56 |
| | $\Sigma$ | 93.71 | 73.58 | **82.43** |

Table 2: System activation performance with ground-truth speech/non-speech segmentation in the DIRHA SimCorpus and DIRHA-L2F RealCorpus test sets.

| Corpus | Set | MissSeg.(%) | Prec. (%) | Recall (%) | F-score (%) |
|---|---|---|---|---|---|
| Sim | test | 10.67 | 67.86 | 50.67 | **58.02** |
| Real | QB | 6.86 | 94.81 | 71.57 | 81.56 |
| | LB | 16.06 | 94.51 | 62.77 | 75.44 |
| | HB | 12.50 | 92.68 | 59.38 | 72.38 |
| | VB | 12.75 | 48.48 | 15.69 | 23.70 |
| | $\Sigma$ | 12.35 | 88.02 | 52.59 | **65.84** |

Table 3: System activation performance with automatic speech/non-speech segmentation in the DIRHA SimCorpus and DIRHA-L2F RealCorpus test sets.

Regarding the overall results using ground-truth segmentation on the two test sets, it is clear that the system performs considerably better in the real data set, even though the parameter tuning was performed on a development simulated data set. The fact of simulating a larger household with 5 rooms, besides including extremely challenging acoustic conditions with several simultaneous overlapping audio events are factors that contribute to this performance difference. In both cases, the Precision score is considerably higher than the Recall. Particularly, in the real test set the overall Precision is remarkably high and it stays more or less constant for the different acoustic conditions. However, the recall clearly decreases for the more challenging acoustic environments. Thus, it seems that the distant noisy and reverberant acoustic environment smooths the speech events in such a way that the probability of matching the background model increases for more challenging acoustic conditions. Consequently, there is an increase of missed system activation detections.

Comparing ground-truth and automatic segmentation results, a generalized performance degradation is observed due to the use of automatic segmentation. Overall results are still better on the real corpus, but the differences between the two tests sets are reduced. In the case of the SimCorpus results, recall decreases around 14%. A considerable contribution to the recall decrease is due to the miss segmentation errors introduced by the SAD module (10.7%). On the other hand, the inserted segmentation errors in combination with the challenging characteristics of the data contribute to the precision performance drop. Regarding the results in the RealCorpus set, an important overall performance drop is observed due to the use of automatic segmentation: F-score from 82.43% to 65.84%. However, a more careful analysis of the real data results by environmental acoustic condition reveals that most of the error contribution is due to the voice-like background noise condition (VB). In these conditions, segmentation errors provoke a drastic drop of both precision (due to an increase of inserted segments) and recall (due to an increase of partially deleted and noise masked segments including activation phrases).

### 4.2. Commands Recognition Evaluation

The overall multi-room and multi-channel voice command recognition processing pipeline WER (%) performance results are shown in Tables 4 and 5 using ground-truth and automatic segmentation respectively. In the latter, all speech segments that actually correspond to a speech event different from a read command are disregarded, while the remaining hypothesized speech segments are processed.

| Corpus | Sets | #Sent. | WER(%) |
|---|---|---|---|
| SimCorpus | test | 75 | **5.15** |
| RealCorpus | QB | 59 | 0.00 |
| | LB | 81 | 1.29 |
| | HB | 40 | 7.72 |
| | VB | 60 | 8.40 |
| | $\Sigma$ | 240 | **3.84** |

Table 4: Command recognition performance using ground-truth speech/non-speech segmentation in the DIRHA SimCorpus and DIRHA-L2F RealCorpus test sets.

Results using ground-truth segmentation (Table 4) for both simulated and real test sets show that the combination of the robust multi-condition acoustic models together with the channel selection strategy allows for extremely good performance in both test sets. The results are particularly remarkable on the real test set taking into account that the system tuning was carried out using simulated data.

Nevertheless, the rather restricted read command recognition grammar used is favoured by the ground-truth segmentation that does not introduce insertions due to wrongly hy-

| Corpus | Sets | Total #Sent. | Match | | +Miss | +Ins | |
|---|---|---|---|---|---|---|---|
| | | | #Sent. | WER(%) | WER(%) | #Sent. | WER(%) |
| Sim | test | 75 | 68 | 13.69 | 21.03 | 2 | **23.27** |
| Real | QB | 59 | 59 | 2.20 | 2.20 | 2 | 3.57 |
| | LB | 81 | 81 | 3.22 | 3.22 | 1 | 3.86 |
| | HB | 40 | 39 | 13.99 | 15.04 | 6 | 23.58 |
| | VB | 60 | 58 | 55.56 | 57.42 | 83 | 165.27 |
| | $\Sigma$ | 240 | 237 | 17.46 | 18.49 | 92 | **47.38** |

Table 5: Command recognition performance with automatic speech/non-speech segmentation in the DIRHA SimCorpus and DIRHA-L2F RealCorpus test sets.

pothesized speech segments (and neither deletions due to lost speech segments). In the case of automatic segmentation results of Table 5, regarding the WER contribution of the detected commands, the performance goes from 5.15% with ideal segmentation to 13.69% in the SimCorpus set, and from 3.84% to 17.46% in the RealCorpus set. The performance drop is more noticeable for more challenging acoustic conditions, particularly for the SimCorpus and the HB and VB RealCorpus sets. Notice that the main reason for this degradation is the inaccurate segmentation of the commands to be processed, which may result in partial deletion of the commands or in commands masked by noise and reverberation embedded in long audio segments. Nevertheless, it is worth noting that, although inaccurate in some acoustic conditions, most of the read commands have been retrieved by the SAD module: more than 90% and 98% are detected in the simulated and real test sets, respectively. Consequently, the impact of missed detected commands in the aggregated WER performance is very limited in the RealCorpus test set (from 17.46% to 18.49%), while it is more relevant in the SimCorpus case (from 13.69% to 21.03%). Finally, regarding the total WER performance that also considers the insertions due to false alarm speech segments, we can observe a large degradation of the RealCorpus results compared to the SimCorpus, mostly due to the poor performance in VB acoustic conditions, i.e., with radio or TV as continuous overlapping interference sources. In the remaining conditions, the proposed multi-channel voice command recognition system achieves remarkable overall voice command recognition performance.

## 5. Conclusions

This paper shows the utility of the DIRHA-L2F RealCorpus for experimental validation of speech processing components in European Portuguese for system activation and voice command recognition in typical multi-room and multi-channel home automation applications.

Excellent results were achieved in the voice command recognition task using ground-truth segmentation information even in the case of extremely challenging acoustic conditions. When using automatic segmentation, the command recognition performance was also remarkable in most of the acoustic conditions. However, it dramatically degraded in the presence of overlapping and continuous voice-like noise, such as TV and radio. In the case of the system activation task, the F-score performance also decreases around 17% with respect to the use of ground-truth segmentation information.

Comparing the processing system performance on real data with respect to simulated data, it can be concluded that results on real data are generally better for both recognition tasks in most conditions. This is partially due to the fact that the simulated corpus characterizes a more challenging home environment of 5 rooms with the regular presence of overlapped speech. Only in the most challenging noise condition of the real corpus, the results are considerably worse. The main source of error for this performance drop are the segmentation errors caused by continuous voice-like noises.

One of the most remarkable characteristics of the complete segmentation, channel selection and speech recognition pipeline is that it is considerably robust to different home environment characteristics. Thus, the system performed remarkably well in real data collected in a totally different acoustic environment to that used for the development of the systems in most of the acoustic conditions. One of the main reasons is that the algorithm considered for multi-channel processing that is exploited both by the segmentation module and the recognition module does not rely in any a priori information or calibration step.

## 6. Acknowledgements

This work was supported by the European Union, under grant agreement FP7-ICT-2011-7-288121, and by funds from the Portuguese Foundation for Science and Technology (FCT) with reference UID/CEC/50021/2013.

## 7. Bibliographical References

Abad, A., Matos, M., Astudillo, R., and Trancoso, I. (2014). The $L^2F$ system for the EVALITA-2014 speech activity detection challenge in domestic environments. In *Proc. Evalita 2014*, Pisa, Italy.

Abad, A. (2007). *A Multi-microphone Approach to Speech Processing in a Smart-room Environment*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Chu, S., Marcheret, E., and Potamianos, G. (2006). Automatic speech recognition and speech activity detection in the CHIL smart room. In *Machine Learning for Multimodal Interaction*. Springer Berlin Heidelberg.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmueller, M., and Maragos, P. (2014). The DIRHA simulated corpus. In *Proc. LREC 2014, Reykjavic, Iceland*.

Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proc. 110th AES Convention*.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The ICSI meeting corpus. In *Proceedings ICASSP'03*.

Jeub, M., Nelke, C., Beaugeant, C., and Vary, P. (2011). Blind estimation of the coherent-to-diffuse energy ratio from noisy speech signals. In *Proc. EUSIPCO 2011*.

Lecouteux, B., Vacher, M., and Portet, F. (2011). Distant speech recognition in a smart home: Comparison of several multisource ASRs in realistic conditions. In *Proc. Interspeech 2011*.

Masgrau, E., Aguilar, L., and Lleida, E. (1999). Performance comparison of several adaptive schemes for microphone array beamforming. In *Proc. EUROSPEECH'99*.

Matos, M., Abad, A., Astudillo, R., and Trancoso, I. (2014). Recognition of distant voice commands for home applications in portuguese. In *Proc. Iberspeech 2014*.

McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., et al. (2005). The AMI meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*.

Meinedo, H. (2008). *Audio pre-processing and speech recognition for Broadcast News*. Ph.D. thesis, Instituto Superior Técnico.

Moreau, N., Mostefa, F., Stiefelhagen, R., Burger, S., and Choukri, K. (2008). Data collection for the chil clear 2007 evaluation campaign. In *Proceedings of LREC'08*, Marrakech, Morocco.

Neto, J. P., Martins, C. A., Meinedo, H., and Almeida, L. B. (1997). The design of a large vocabulary speech corpus for Portuguese. In *Proc. Eurospeech'97*.

Obuchi, Y. (2004). Multiple-microphone robust speech recognition using decoder-based channel selection. In *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*.

Ravanelli, M., Sosi, A., Svaizer, P., and Omologo, M. (2012). Impulse response estimation for robust speech recognition in a reverberant environment. In *Proceedings EUSIPCO 2012*.

Seltzer, M. L. (2003). *Microphone array processing for robust speech recognition*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA.

Tsiami, A., Rodomagoulakis, I., Giannoulis, P., Katsamanis, A., Potamiannos, G., and Maragos, P. (2014). ATHENA: A Greek Multi-Sensory Database for Home Automation Control. In *Proc. Interspeech 2014*.

Vacher, M., Lecouteux, B., Chahuara, P., Portet, F., Meillon, B., and Bonnefond, N. (2014). The sweet-home speech and multimodal corpus for home automation interaction. In *Proc. LREC 2014, Reykjavic, Iceland*.

Wolf, M. and Nadeu, C. (2010). On the potential of channel selection for recognition of reverberated speech with multiple microphones. In *Proc. Interspeech 2010*.

Wolf, M. (2013). *Channel Selection and Reverberation-Robust Automatic Speech Recognition*. PhD, Universitat Politècnica de Catalunya (UPC).

Wölfel, M., Fügen, C., Ikbal, S., and McDonough, J. (2006). Multi-source far-distance microphone selection and combination for automatic transcription of lectures. In *Proc. Interspeech 2006*.