10.12871/clicit2014227

# The L$^2$F system for the EVALITA-2014 speech activity detection challenge in domestic environments

**Alberto Abad, Miguel Matos, Hugo Meinedo, Ramon F. Astudillo, Isabel Trancoso**
INESC-ID/IST Lisbon, Portugal
{alberto.abad, jmatos, hugo.meinedo, ramon.astudillo, isabel.trancoso}@l2f.inesc-id.pt

## Abstract

**English.** The INESC-ID's Spoken Language Systems Laboratory (L$^2$F) submission to EVALITA-2014 targets the problem of room-localized speech activity detection in multi-room domestic environments. The three proposed systems, which have been developed within the activities of the DIRHA project, combine multi-channel model-based speech classification with automatic room localization, based on spectral envelope distortion measures. The processing chain of the investigated approaches is composed of three basic stages: 1) multi-channel speech segmentation is carried out for each room, 2) speech segments detected at each room are time-aligned, and 3) a room assignment strategy is applied to each candidate speech event to determine in which room it was generated. The three submitted systems exploit the same speech/non-speech adapted model and the same channel combination strategy, while differing in the room localization strategy. Results obtained in the official EVALITA-2014 task confirm the effectiveness of the proposed methods. Particularly, in the case of real test data, F-scores of 98.1% are attained.

**Italiano.** *Il sistema sottomesso da INESC-ID Spoken Language Systems Laboratory (L$^2$F) affronta il probelma del rilevamento del parlato con relativa assegnazione ad una stanza in un tipico ambiente domestico caratterizzato da numerose stanze. I tre sistemi proposti, sviluppati nell'ambito del progetto DIRHA, combinano una prima classificazione del parlato, ottenuta attraverso un'elaborazione multi canale, con una selezione della stanza basata sulla distorsione dell'inviluppo spettrale. Il sistema e' costituito da tre componenti: 1) una segmentazione multi canale effettuata su ogni stanza; 2) i segmenti identificati sono allineati temporalemente; 3) una stanza viene assegnata ad ogni candidato. I tre sistemi adottano lo stesso modello di speech/non-speech e la stessa strategia nel combinare i canali, mentre si differenziano nel modo in cui viene selezionata la stanza da associare a ciascun evento. I risultati ottenuti sul task ufficiale di EVALITA-2014 confermano la convenienza dei metodi presentati. In particolare, sui dati reali i sistemi proposti raggiungo una F-score pari al 98.1%.*

## 1 Introduction

Speech activity detection of the acoustic input constitutes a crucial component in any voice-enabled application, providing important information to other system components, such as speaker localization, keyword spotting, automatic speech recognition, and speaker recognition, among others. In general, the quality of the segmentation information has a huge impact on the following speech processing components and its relevance is exacerbated for services that are required to work in an "always-listening" mode. This is the case of home automation applications. In fact, for such domestic scenarios, additional challenges affecting the performance of speech activity detection usually arise. First, microphones are normally located far from the source speaker in an environment that can be highly dynamic, noisy and reverberant. Second, in addition to detect "when" a speech activity has taken place, in multi-room environments it is important to decide "where" in the house such activity occurred.

The Speech Activity detection and Speaker


Figure 1: Block diagram of the speech/non-speech segmentation module.

LOcalization in DOMestic environments (SASLODOM) challenge, that is part of the EVALITA'2014 evaluation campaign, focuses on the detection and localization of speech events generated by users within the various rooms of a household. The scenario addressed in the task is the one of the DIRHA project (DIRHA, 2012), that is, an apartment monitored by 40 microphones, distributed on the walls and the ceiling of its five rooms. It encompasses typical situations observable in domestic contexts, in terms of speech input as well as of other acoustic events and background noise. For each speech event, the goal of the task is to: a) provide the corresponding time boundaries, b) determine the room where it was generated, and c) derive the spatial coordinates of the speaker. The task is evaluated in both simulated and real data sets in Italian, created by the DIRHA consortium. Additional details about the task, including guidelines, data, evaluation tools, details about the rooms and about the microphones are available in the SASLODOM task report (A. Brutti et al, 2014).

This report describes the L$^2$F speech activity detection (SAD) systems submitted to the SASLODOM challenge. The proposed systems have been developed within the activities of the DIRHA project.The complete room-localized SAD system is based on a three stage process. First, multi-channel speech segmentation is carried out for each room. Second, speech segments detected at each room are time-aligned in order to identify speech events that are likely to be the same. Third, a room assignment strategy is applied to each candidate speech event to determine in which room it was generated.

## 2 The L$^2$F multi-room SAD systems for domestic environments

The L$^2$F multi-room SAD systems have been developed in the context of the DIRHA project. This section provides details on different approaches investigated and evaluated using DIRHA data.

### 2.1 The DIRHA SimCorpus

The DIRHA SimCorpus (L. 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) (M. Ravanelli et al, 2014). In this work, the development set of the DIRHA SimCorpus has been used to adapt the speech/non-speech model that is part of the SAD module (more details in section 2.2). On the other hand, the test set of the European Portuguese DIRHA SimCorpus is used to assess the different methods under study.

### 2.2 Baseline MLP-based SAD detector

The core module of the L$^2$F systems is a model-based speech/non-speech classifier. This module is composed by several blocks, as depicted in Figure 1. The first one, designated as feature extraction, performs acoustic parametrization of the audio signal, extracting 12th order perceptual linear prediction (PLP) coefficients plus signal frame energy, all appended by their first temporal derivatives, thus yielding 26-dimensional acoustic features. These are subsequently passed to the classification block, which is implemented using an artificial neural network of the multi-layer perceptron (MLP) type (Meinedo, 2008). The baseline neural classifier was trained using 50 hours of TV Broadcast News and 41 hours of varied music and sound effects (in order to improve the representation of non- speech audio signals). The output of the trained neural classifier represents the probability of the audio signal containing speech. The following block smooths this probability using a median filter over a small window. The smoothed signal is then thresholded and analysed using a time window ($t_{min}$). The final block is a finite state machine that consists of four possible states ("probable non-speech" "non-speech", "probable speech", and "speech"). More details can be found in (A. Abad et al, 2013).

## 3 Baseline for distant speech recognition in Portuguese

### 3.1 Improvements to the baseline SAD

The aim of this section is to improve the baseline SAD module. For that purpose, we define a new task that consists of detecting speech events occurring in a specific room and ignoring the speech events that occur in the other rooms. We refer to this task as the "isolated-room" SAD task. Notice that this is not the targeted task in the SASLODOM challenge. Nevertheless, this "isolated-room" SAD task permits the assessment of the proposed systems ignoring the errors due to cross-room speech insertions, which is a particularity of multi-room environments. In this section, the DIRHA SimCorpus for European Portuguese (PT) was used for testing.

#### 3.1.1 MLP adaptation

The MLP model described previously is not at all adjusted to the acoustic environments targeted at DIRHA. A reasonable solution for this problem is to retrain or adapt the MLP based classifier using appropriate data, that is, data more similar to the test conditions. To evaluate the feasibility of this approach, the baseline MLP classifier was adapted using three development sets from the DIRHA SimCorpus, namely the ones in Italian (IT), European Portuguese (PT), and Greek (GR). As described in (M. Ravanelli et al, 2014), the simulated data correspond to microphones located in five rooms of the apartment. For each room, a specific microphone was chosen. A total of 1125 audio files from the 3 languages, 5 rooms, and 75 recorded simulations were used in the adaptation, of which 750 for training and the remaining 375 to validate the model. The MLP was fully adapted using a single epoch of back-propagation, with a much smaller learning step than the one used for the initial model training.

#### 3.1.2 Multi-channel combination

In addition to the adaptation of the speech/non-speech model, improved segmentation for each room is obtained by exploiting all the microphones available in the apartment. We explore two methods of multi-channel combination: Majority Voting Decision Fusion (MVF) and Posterior Probability Fusion (PF).

**Majority Voting Decision Fusion (MVF)** In the MVF method, the baseline speech/non-speech segmentation module is first run individually for each channel of the house. Then, the resulting segmentations from all the channels of a specific room are aligned to detect candidate speech events. Due to the possible different propagation delays from the speech source to the several microphones, a tolerance of 1 second is given to this alignment process. Then, if more than half of the microphones of a specific room detect a speech event candidate, the system considers that there was speech in that room in that time interval.

**Posterior Probability Fusion (PF)** In the PF method, the posterior probabilities obtained by the MLP classifier for each channel of a specific room are combined before applying the median filter. The combination rule is simply the mean of the probabilities provided by the MLP. Then, the same finite state machine adopted in the single-channel case is used to obtain the room segmentation based on these averaged probabilities.

#### 3.1.3 "Isolated-room" SAD task results

The results of the distinct approaches are presented in Table 1. In the mono-channel system, a representative microphone was chosen for each room. Observing the speech recall values of Table 1, it can be seen that the MLP unadapted system (*MLP-Baseline*) rejects a very high percentage of speech. After adaptation of the network classifier with in-domain data (*MLP-DIRHA*), speech recall increases to around 80%, while maintaining a high non-speech detection precision. Regarding multi-channel combination approaches, generalized improvements (F-score) are attained with respect to the mono-channel approach. There are no significant differences between the two multi-channel methods.

### 3.2 Room-Localized SAD

In this section, we focus on the SASLODOM task, that we refer to as "room-localized" SAD task. Notice that in contrast to the previous section, the detected speech segments which originated in other rooms are considered as insertion errors and affect the performance of the evaluated systems. Table 2 presents the results achieved by the SAD systems previously described when evaluated in the "room-localized" task. As it can be observed, performances greatly decrease compared to the ones reported in Table 1. This is due to the high rate of detected speech segments actually occur-

| System [channel + MLP model] | speech | | | non-speech | | | total |
|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F-score | Prec. | Recall | F-score | Acc. |
| 1c + MLP-Baseline | 99.7 | 54.7 | 70.6 | 95.2 | 100 | 97,5 | 95.4 |
| 1c + MLP-DIRHA | 70.8 | 81.0 | 75.5 | 97.8 | 96.3 | 97.0 | 94.7 |
| MVF + MLP-DIRHA | 74.2 | 80.7 | 77.3 | 97.8 | 96.9 | 97.3 | 95.2 |
| PF + MLP-DIRHA | 76.1 | 79.9 | 77.9 | 97.7 | 97.2 | 97.5 | 95.5 |

Table 1: Performance (%) of the "isolated-room" speech activity detection task with the European Portuguese DIRHA SimCorpus test set using different MLP classifiers with single-channel and multi-channel combination approaches.

| System [channel + MLP model] | speech | | | non-speech | | | total |
|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F-score | Prec. | Recall | F-score | Acc. |
| 1c + MLP-DIRHA | 26.1 | 81.6 | 39,5 | 98.2 | 81.1 | 88.8 | 81.1 |
| MVF + MLP-DIRHA | 26.5 | 81.4 | 40.0 | 98.2 | 81.5 | 89.1 | 82.5 |
| PF + MLP-DIRHA | 27.5 | 80.4 | 41.0 | 98.1 | 82.7 | 89.7 | 81.5 |

Table 2: Performance (%) of the "room-localized" speech activity detection task with the European Portuguese DIRHA SimCorpus test set using different MLP classifiers with single-channel and multi-channel combination approaches.

ring in a different room. These results show the inadequacy of the proposed approaches for the targeted task.

#### 3.2.1 Strategies for room detection

In order to address the cross-room detection problem, we propose to combine conventional SAD approaches with automatic room detection methods. The proposed method consists of a three-step process as follows:

1. Obtain automatic segmentation for each room using any of the previously described methods. With this operation, we obtain a set of speech candidate segments for each room.

2. Align speech candidate segments of all rooms with a tolerance of 1 second. This is done to match events that are likely to be the same ones, but that are simultaneously detected at different rooms.

3. Decide to which room every speech candidate segment belongs using the information provided by an automatic room detector.

From the various room-detection methods studied, the ones based on envelope variance (EV) distortion measures (M. Wolf and C. Nadeu, 2010) were chosen, because they present the best trade-off between computational load and performance for an environment with noise and reverberation. In this work, the detected room corresponds to the room of the microphone with the highest EV measure in the time interval of the candidate speech segments. In practice, we have explored two methods of integrating the segmentation information and the room localization information:

- *Restricted room selection (Restricted-RS)* The rooms in which the speech event may happen are restricted to those rooms that actually detected that hypothesised segment.

- *Matched room selection (Matched-RS)* Automatic room detection is not restricted and any room may be selected for each hypothesised speech segment. However, if the automatically selected room does not match any of the rooms that actually detected the hypothesized segment, then that candidate segment is disregarded.

In practice, the difference between the two methods is that in the first case, all aligned candidate segments are assigned to one room (and removed from any other room in which the same candidate is detected), while in the second case, there may be candidate segments that are disregarded and not assigned to any room. Consequently, for the second approach, one may expect an increase of the precision in exchange for a drop in the recall performance.

| Room selec. approaches | System [channel + MLP model] | speech | | | non-speech | | | total |
|---|---|---|---|---|---|---|---|---|
| | | Prec. | Recall | F-score | Prec. | Recall | F-score | Acc. |
| *Restricted-RS* | 1c + MLP-DIRHA | 43.2 | 65.9 | 52.2 | 97.1 | 92.9 | 95.0 | 90.9 |
| | MVF + MLP-DIRHA | 46.4 | 65.3 | 54.3 | 97.1 | 93.8 | 95.4 | 91.7 |
| | PF + MLP-DIRHA | 46.9 | 65.6 | 54.7 | 97.1 | 93.9 | 95.5 | 91.8 |
| *Matched-RS* | 1c + MLP-DIRHA | 73.2 | 59.5 | 65.7 | 96.7 | 98.2 | 97.5 | 95.3 |
| | MVF + MLP-DIRHA | 75.2 | 59.6 | 66.5 | 96.7 | 98.4 | 97.6 | 95.5 |
| | PF + MLP-DIRHA | 74.9 | 59.8 | 66.5 | 96.8 | 98.4 | 97.6 | 95.4 |

Table 3: Performance (%) of the "room-localized" speech activity detection task with the European Portuguese DIRHA SimCorpus test set using applying single-channel and multi-channel fusion approaches combined with two different room-localization approaches based in EV.

| Test data | System [channel + MLP model + RS] | O-SAD | FA | DR | Prec. | Recall | F-score |
|---|---|---|---|---|---|---|---|
| *Simulated* | MVF + MLP-DIRHA + Non-RS | 7.7 | 12.0 | 3.4 | 53.5 | 95.9 | 68.7 |
| | MVF + MLP-DIRHA + Restricted-RS | 11.8 | 5.4 | 18.3 | 73.4 | 79.2 | 76.2 |
| | MVF + MLP-DIRHA + Matched-RS | 14.4 | 3.6 | 25.2 | 82.3 | 75.1 | 78.5 |
| *Real* | MVF + MLP-DIRHA + Non-RS | 13.7 | 26.1 | 1.3 | 49.2 | 96.2 | 65.1 |
| | MVF + MLP-DIRHA + Restricted-RS | 2.0 | 2.7 | 1.3 | 100 | 96.2 | 98.1 |
| | MVF + MLP-DIRHA + Matched-RS | 2.0 | 2.7 | 1.3 | 100 | 96.2 | 98.1 |

Table 4: Performance results (%) of the $L^2F$ speech activity detection systems submitted to the SASLODOM challenge in the simulated and real data test sets in terms of the official task evaluation metrics: Overall SAD performance (O-SAD), false alarm rate (FA), deletion rate (DR), Precision (Prec), Recall and F-score.

#### 3.2.2 "Room-Localized" SAD task results

Table 3 presents the results obtained for the two integrated approaches that combine speech activity detection and room localization. Comparing these results with the ones obtained with the systems that do not incorporate any room assignment strategy (Table 2), we can observe a great improvement in the precision performance of speech. On the other hand, there is also a considerable drop in the recall performance. However, we can see that the incorporation of room localization increases the system performance about 25% for the best method in terms of F-score. These results seem to demonstrate the convenience of the methods proposed that combine segmentation with room localization.

Regarding the room-assignment strategies, the recall is higher for the *Restricted-RS* approach, as expected, because all candidate segments are always assigned to one room. On the other hand, also as expected, the precision is very low when compared to the *Matched-RS* approach. In general, the second approach achieves a better generalised performance (F-score).

## 4 The $L^2F$ SASLODOM 2014 submission

Three different systems have been submitted to the EVALITA-SASLODOM 2014 challenge. The three systems differ in the room selection strategy integrated: no room selection (*Non-RS*), restricted room selection (*Restricted-RS*) and matched room selection (*Matched-RS*). The three systems share the same MLP classifier adapted with in-domain data (MLP-DIRHA), since it showed remarkable improvements with respect to the baseline classifier in the experiments with the DIRHA SimCorpus. Moreover, given that no significant performance differences were observed regarding multi-channel combination methods, majority voting fusion (MVF) approach was applied in all cases. It is worth noting that system tuning has not been conducted to adapt to the particular characteristics of the SASLODOM data.

Table 4 shows the official performance results obtained by the submitted systems in the simulated and real data test sets. According to these results, the trends of the different systems are as expected: the highest recall/lowest precision is achieved by the system that does not incorporate

room detection strategies, while the *Matched-RS* is the room assignment strategy that provides highest precision in exchange for a moderate recall drop. Regarding F-score metrics, the *Matched-RS* approach is the best performing one. Comparing the *Simulated* results to the ones reported in the previous section, two relevant differences can be noticed. First, the general performance is considerably better: F-scores increase from 40.0%, 54.3% and 66.5% to 68.7%, 76.2% and 78.5%, for each of the three submitted systems respectively. Second, the performance differences between the three systems are considerably reduced. A possible explanation for these two observations may be the reduced amount of cross-room detected speech events in the SASLODOM data when compared to the DIRHA data. However, this is only an hypothesis that needs to be further investigated and there may be other explanations for the observed phenomena. Finally, it is worth highlighting the extremely good performances with real data (F-score 98.1%) achieved by the proposed approaches incorporating automatic room detection information. Note that these methods allowed for a drastic precision increase, from 49.2% to 100%, while keeping the recall constant at 96.2%. These figures show that each candidate speech segment is in fact simultaneously detected at the two rooms. However, the room assignment strategy based on EV is able to perfectly determine the correct room where each speech event is generated. This result confirms the effectiveness of the EV distortion metric for channel and room selection with real data.

## Acknowledgements

This work was partially supported by the European Union, under grant agreement FP7-ICT-2011-7-288121, and by the Portuguese Foundation for Science and Technology, through project PEst-OE/EEI/LA0021/2013 and grant number SFRH/BPD/68428/2010. The authors would like to thank to their colleagues in the DIRHA consortium and to the organizers of the EVALITA-SASLODOM 2014 challenge.

## References

DIRHA project. 2012. `http://dirha.fbk.eu/`.

A. Brutti et al. 2014. "SASLODOM: Speech Actvity detection and Speaker LOcalization in DOMestic environments," in *Proceedings of Evalita 2014.*. Pisa University Press, 2014.

L. Cristoforetti et al. 2014. "The DIRHA simulated corpus," in *Proc. LREC 2014*

M. Ravanelli et al. 2014. "DIRHA-simcorpora I and II," *Deliverables 2.1, 2.3, 2.4, DIRHA Consortium.*

H. Meinedo. 2008. "Audio pre-processing and speech recognition for BroadcastNews," Ph.D. dissertation, IST, Lisbon, Portugal.

A. Abad et al. 2013. "Multi-microphone front-end," *Deliverable D3.2, DIRHA Consortium*.

M. Wolf and C. Nadeu. 2010. "On the potential of channel selection for recognition of reverberated speech with multiple microphones," in *Proc. Interspeech 2010*:80–83.