The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones

Marco Matassoni, Ramón Fernandez Astudillo, Athanasios Katsamanis, Mirco Ravanelli

1Fondazione Bruno Kessler (Italy)
2L2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
3National Technical University of Athens (Greece)
{matasso,mravanelli}@fbk.eu, ramon.astudillo@inesc-id.pt, nkatsam@cs.ntua.gr

Abstract

Distant speech recognition in real-world environments is still a challenging problem and a particularly interesting topic is the investigation of multi-channel processing in case of distributed microphones in home environments. This paper presents an initiative oriented to address the challenges of such a scenario: an experimental recognition framework comprising a multi-room, multi-channel corpus and the accompanying evaluation tools is made publicly available. The overall goal is to represent a common platform for comparing state-of-the-art algorithms, share ideas of different research communities and integrate several components in a realistic distant-talking recognition chain, e.g., voice activity detection, speech/feature enhancement, channel selection and fusion, model compensation. The recordings include spoken commands (derived from the well-known GRID corpus) mixed with other acoustic events occurring in different rooms of a real apartment. The work provides a detailed description of data, tasks and baseline results, discussing the potential and limits of the approach and highlighting the impact of single modules on recognition performance.

Index Terms: multi-channel processing, distant speech recognition, voice activity detection, speech enhancement, feature compensation, model adaptation

1. Introduction

Applications based on distant speech recognition [1] in real-world environments still suffer from inadequate robustness; reverberation and dynamic background noise represent major sources of acoustic mismatch that heavily decrease automatic speech recognition (ASR) performance, which, on the contrary, can be very good in close-talking microphone setups [2, 3, 4]. In this context, a particularly interesting topic is the adoption of distributed microphones for the development of automated home environments controlled via distant-speech interaction. Nowadays, automated domotic devices are typically controlled from the house, have to be filtered out and rejected.

In this regard and in the frame of DIRHA, specific acoustic corpora are being collected or created [7] in an effort to reproduce the typical acoustic phenomena characterizing the scenario in focus and facilitate research and development activities specifically addressing the involved issues. So far, four different corpora have been collected and made available in four different languages, namely German, Greek, Italian and Portuguese, and are currently actively used in speech recognition, keyword spotting, source localization and acoustic event detection experiments, as, e.g., in [8, 9]. Following these developments and exploiting the already available and popular GRID corpus [10], in this paper we introduce the DIRHA-GRID corpus, which is in English, and the corresponding baseline that aim to serve as additional, useful and more accessible resources for the communities working on these topics.

The accompanying task, namely distant speech recognition of multiple commands intermixed with other acoustic events in one-minute long recordings, represents an attempt to define a single task requiring the combination of all the components that significantly more natural interaction. Indeed, speech recognition is progressively entering in this application field. However, unresolved robustness and flexibility issues related to the typical usage conditions, e.g., spontaneous speech, hands-free interaction and uncontrolled acoustic conditions, currently hinder this development.

The variabilities related to microphone location and unpredictable acoustic interferences are critical problems and the scientific research is being addressing these issues in order to enable far-field speech recognition even in the presence of other acoustic sources. Several projects dealt with recognition in meeting rooms [5, 6] where the various challenges of distant speech recognition are addressed: localization, detection, segmentation, enhancement and acoustic model training or adaptation. These problems are currently under investigation in the DIRHA project [1], where realistic home environments are studied. The underlying vision is to develop an automated system based on a number of microphones installed in different rooms of a real apartment; the system will selectively monitor the speech activities in the household, detect and understand the voice commands, and act as the interface for the full control of appliances and devices. In this direction, the coherent processing of multi-microphone signals to detect and properly manage concurrent acoustic events in different rooms becomes of critical importance; conversely, all the possible interfering sources, e.g., phone ringing, radio, background noise external to the house, have to be filtered out and rejected.

The research leading to these results has partially received funding from the European Union’s 7th Framework Programme (FP7/2007-2013) under grant agreement n. 288121 - DIRHA. R. F. Astudillo’s work was funded by the FCT through the grant number SFRH/BPD/88428/2010 and project PEst-OE/EEI/LA0021/2013. We would like to thank ITEA S.p.A (Istituto trentino per l’edilizia abitativa) to have kindly made available their automated apartment for experiments under the DIRHA project. We also thank Jon Barker and colleagues for making available the clean corpus and the related recognition setup.

1http://dirha.fbk.eu
need to be employed in a realistic distant-talking recognition system. Along with data, tools to perform ASR and solve some of the sub-problems, e.g., voice activity detection, feature and model compensation, are provided, and have already become available to interested research groups (see the corresponding website \(^2\)).

The paper is organized as follows: Section 2 introduces the motivation and the challenges of the proposed experimental framework. Section 3 describes the corpus released for this investigation and the recognition task while the accompanying tools are listed in Section 3.1. The experimental baseline is then introduced in Section 4 and the results are discussed in Section 4.4. Finally, Section 5 presents our concluding remarks.

2. Motivation and challenges

The goal of the paper is to investigate multi-channel processing for ASR in domestic environments, where microphones are installed in different rooms and the resulting multi-channel audio recordings capture multiple audio events, including voice commands or spontaneous speech, generated in various locations and characterized by a variable amount of reverberation as well as possible background noise \([11]\).

The adoption of array-processing techniques is not always possible or may be problematic \([12]\) and alternative approaches can be successfully applied, as for example channel selection \([13]\) or source separation \([14, 15]\). Indeed, the availability of multi-channel data enables a variety of approaches such as spatial filtering, interference cancellation, post-filtering \([16]\). Spatial processing can exploit the information available about the desired acoustic source, the geometry of sound propagation, and the competing or background noise sources. Algorithms based on blind source separation can complement the spatial processing and improve the overall performance. Another important component is represented by speech activity detection: the system usually runs in always-listening mode to satisfy the hands-free concept so it becomes crucial to correctly estimate the time boundaries of the speech events to be recognized. Besides speech detection, it is often necessary to apply some form of enhancement or selection on the acquired signals. Indeed, the microphone channels are characterized by different acoustic phenomena that occur in the different rooms and the recognizer has to be fed with the best signal \([8]\). Other topics to be considered are feature enhancement \([18, 19]\) or compensation \([20]\) and acoustic modeling \([21, 22]\). Hence, the complexity of the investigated scenario makes the design of a realistic acoustic corpus difficult. Some of the earlier initiatives \([14, 23, 24]\) have successfully addressed a number of typical problems encountered in similar environments and inspired this proposal, in particular the choice of GRID \([10]\) as the source for the dry signals. The key motivations are:

- creation of a small-vocabulary speech recognition task based on publicly available resources and characterized by highly confusable tokens (e.g., letters);
- exploration of the acoustic variability offered by multiple microphones distributed in different rooms observing simultaneously complex acoustic scenes;
- incorporation of some typical problems often not considered: voice activity detection, selection or fusion of redundant information, etc.;

\(^2\)http://dirha.fbk.eu/interspeech2014

- evaluation of the performance using a simple recipe and possibility to compare the investigated techniques with oracle results.

Some characteristics have been directly inherited from the pioneering work of the CHiME challenges to make this new initiative more accessible to people who have already participated in these challenges. All these elements have motivated the design of the corpus described in the Section 3 to address new specific problems, namely multi-room recordings with distributed microphones and long sequences with multiple commands. Of course, the presented corpus, despite being quite realistic, is still limited compared to a real data collection. However, it provides a controlled testbed for research and experimentation and corresponding results, as presented in Section 4, can demonstrate the real challenges of the investigated scenario.

3. The multi-channel corpus

The DIRHA-GRID corpus is a multi-microphone and multi-room simulated database developed under the DIRHA project. The corpus contains a set of acoustic scenes of 1-minute duration (at 16kHz sampling frequency and 16-bit accuracy) which are observed by 40 microphone channels distributed over five different rooms (15 in the living-room, 13 in the kitchen, 7 in the bedroom, 3 in the bathroom and 2 in the corridor) of a real apartment. The overall multi-microphone setup is depicted in Figure 1.

![Figure 1: An outline of the microphone set-up. Black dots represent microphones, while boxes and arrows indicate speaker positions and orientations.](image)

Each acoustic scene is composed of both speech, i.e., short English commands from the GRID database \([10]\), and non-speech acoustic sources, i.e., typical home noises. Each acoustic event (both speech and non-speech) occurs randomly in time and in space and can take place in any of the microphone-equipped rooms. In particular, a variable number of short commands (ranging from 4 to 7) arise in each 1-minute long acoustic scene. An overlap in time between speech and non-speech sources is possible, while an overlap between speech sources cannot occur. The corpus is divided into 3 chunks (dev1, test1, test2) containing 75 acoustic sequences each with 12 different speakers (6 male and 6 female) involved for each dataset. The development set (dev1) is intended to be adopted to tune the pa-
rameters of the proposed system while the other sets (test1 and test2) are released for testing purposes. An XML-based annotation describing in detail each sequence and a picture which depicts each acoustic scene complete the documentation. This corpus is partially aligned with the multi-microphone and multi-language DIRHA-simcorpora [7], which have been developed under the project and already partially released [25].

The overall multi-microphone set-up (Figure 1) is based on both distributed microphone networks (pairs or triplets of sensors on the walls) and more compact arrays composed of 6 microphones (on the ceiling of the living-room and the kitchen). In each pair, the sensors are at a distance of 30 cm, while the arrays are placed on a circumference of radius 30 cm with one sensor arranged at the center of the circle.

Each 1-minute acoustic scene includes acoustic events occurring at different possible time-instants, rooms and positions: a) speech and other acoustic events that occur in the house. For the baseline, this is limited to Voice Activity Detection (VAD). It has to be taken into account that in our setup AED plays a more fundamental role compared to robust ASR corpora as, e.g., the AURORA tasks [28] or the REVERB task [24]. The DIRHA-GRID task features one minute-long audio simulations with multiple commands uttered from different locations. An optimal processing of the simulation requires therefore first locating each event within the file to eliminate background frames that induce errors and second identify the best microphones to process that event. For the proposed baseline, a simple multichannel model-based VAD system is trained on the development set. Specifically, two separate Gaussian Mixture Models (GMMs) are trained for speech and background respectively and activity detection is achieved by means of the Viterbi algorithm, considering two different states. This allows the imposition of a penalty when switching from the one state to the other which results in fewer spurious event detections. By allowing for a class-specific penalty one can properly tune the classifier to increase recall of speech at the expense of low precision. Excessive false alarms are then expected to be rejected by the following ASR module. In the case of the AFE, its frame-dropping utility was used when no oracle information was available.

**Microphone-Network Processing:** this concerns the identification of one or more microphones as suitable to process each individual speech event and the processing scheme used. No specific methods were however developed for the baseline. The oracle information available in the DIRHA-GRID corpus was used instead, as detailed in Section 4.2, to identify either the room were each speech activity occurred or the speaker position. In the former case, the central microphone of each room, e.g., B1L, LA6, KA3, R1R, C1R, is used for recognition. In the latter, a Delay and Sum (DS) beamformer over all room microphones was implemented. For the AFE, the microphone closest to the speaker was chosen, as it led to better results. When no oracle information was available the central microphone of the living room LA6 was used.

**Single-Channel Enhancement:** two sub-modules were considered. The noise estimation block carried out an estimation of a priori signal to noise ratio and noise variance. This estimates were then used to produce an estimate of the clean signal in the speech enhancement sub-module. Two strategies were tested for the baseline: using the background frames eliminated at the AED stage to estimate noise by recursive smoothing and using the IMCRA SNR estimator [29], in addition to the latter, to update the noise estimate during speech activity. When IMCRA was used, preceding background frames were used to initialise the algorithm for each event. To increase the performance of the approach, propagation of the posterior associated with the Wiener filter through an intermediate resynthesis step was used. This utilized the ISTFT+STFT Uncertainty Propagation (IUP) approximation introduced in [30]. The approximation was nevertheless used to propagate the posterior, rather than the prior distributions.

### 4. Baseline

To facilitate evaluation of the proposed algorithms, we have provided tools to perform ASR and solve some of the sub-problems of the task such as acoustic event detection, feature and model compensation. To allow participants to concentrate on the sub-problems that they prefer, various tools, provided in modular form, are hence available. These include a) an HTK recipe for training speaker independent models from the clean GRID database; this also includes supervised MLLR adaptation to the development set; b) HTK recipe for decoding and WER evaluation on the DIRHA-GRID signals using a pre-defined grammar (repetitions of commands plus optional silence); c) HTK Voice Activity Detector based on GMMs and adapted to the development set. This can be used to create a transcription (MLF) file to extract speech from the audio files; d) HTK recipe-integrated Matlab tools to extract acoustic features, including default MFCC feature extraction and supporting frame-dropping. Finally, we also provide Matlab tools and HTK patches for feature and model compensation. These include a MMSE-MFCC estimator [26] and HTK patches for Uncertainty Decoding and Modified Imputation.

### 4.1. Modules Considered

**Acoustic Event Detection:** this concerns the detection of speech and other acoustic events that occur in the house. For this concerns the identification of one or more microphones as suitable to process each individual speech event and the processing scheme used. No specific methods were however developed for the baseline. The oracle information available in the DIRHA-GRID corpus was used instead, as detailed in Section 4.2, to identify either the room were each speech activity occurred or the speaker position. In the former case, the central microphone of each room, e.g., B1L, LA6, KA3, R1R, C1R, is used for recognition. In the latter, a Delay and Sum (DS) beamformer over all room microphones was implemented. For the AFE, the microphone closest to the speaker was chosen, as it led to better results. When no oracle information was available the central microphone of the living room LA6 was used.

**Microphone-Network Processing:** this concerns the identification of one or more microphones as suitable to process each individual speech event and the processing scheme used. No specific methods were however developed for the baseline. The oracle information available in the DIRHA-GRID corpus was used instead, as detailed in Section 4.2, to identify either the room were each speech activity occurred or the speaker position. In the former case, the central microphone of each room, e.g., B1L, LA6, KA3, R1R, C1R, is used for recognition. In the latter, a Delay and Sum (DS) beamformer over all room microphones was implemented. For the AFE, the microphone closest to the speaker was chosen, as it led to better results. When no oracle information was available the central microphone of the living room LA6 was used.

**Single-Channel Enhancement:** two sub-modules were considered. The noise estimation block carried out an estimation of a priori signal to noise ratio and noise variance. This estimates were then used to produce an estimate of the clean signal in the speech enhancement sub-module. Two strategies were tested for the baseline: using the background frames eliminated at the AED stage to estimate noise by recursive smoothing and using the IMCRA SNR estimator [29], in addition to the latter, to update the noise estimate during speech activity. When IMCRA was used, preceding background frames were used to initialise the algorithm for each event. To increase the performance of the approach, propagation of the posterior associated with the Wiener filter through an intermediate resynthesis step was used. This utilized the ISTFT+STFT Uncertainty Propagation (IUP) approximation introduced in [30]. The approximation was nevertheless used to propagate the posterior, rather than the prior distributions.

### 4.2. Oracle Knowledge Used

<table>
<thead>
<tr>
<th>Oracle</th>
<th>AED</th>
<th>MNP per Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>O1</td>
<td>Loose VAD Oracle</td>
<td>None</td>
</tr>
<tr>
<td>O2</td>
<td>Loose VAD Oracle</td>
<td>Room Oracle</td>
</tr>
<tr>
<td>O3</td>
<td>Loose VAD Oracle</td>
<td>Location Oracle</td>
</tr>
</tbody>
</table>

Table 1: Different oracle information used.
Efficiency in performance.
The combination of IMCRA with IUP also had an important
ing speech events to improve the estimate of background noise.
This error can be brought down by using IMCRA dur-
background frames provided by the AED and a MMSE-LSA es-

The largest improvement is obtained estimating noise from the
speaker is used rather than DS for comparability with the AFE.

Table 3 details the im-
provements attained with each speech enhancement sub-module
under O3 conditions. Note that the microphone closest to
speaker is used rather than DS for comparability with the AFE.
The largest improvement is obtained estimating noise from the
background frames provided by the AED and a MMSE-LSA es-
timator. This error can be brought down by using IMCRA dur-
ing speech events to improve the estimate of background noise.
The combination of IMCRA with IUP also had an important
effect in performance.

4.3. Training and Parameter Adaptation
The training and adaptation followed the HTK recipe released
along with the DIRHA-GRID data. This recipe showed how-
ever overfitting problems, as discussed in the results section.

The recipe employed the original GRID corpus to train clean
models with a simple front-end. Specific models were trained
using MFCC features with no processing for the baseline and
the AFE. For a given processing scheme, including AED and
MNP modules, the MLLR recipe provided in the DIRHA-GRID
task was used.

Regarding speech enhancement, the IMCRA estimator and
the MMSE clean speech estimators were used with their orig-
inal configuration. However, the design of the approach was
carried out on the development set. This implied the selection
of the best MMSE estimator, in this case the MMSE-LSA, and
the use of the posterior propagation using IUP as optimal meth-
ods. Other alternatives as using uncertainty of observation tech-
niques were explored but excluded from the final baseline.

4.4. Results
Effect of each module on performance: Table 2 shows the re-
results of the proposed baseline with and without single-channel
processing and the AFE for different levels of oracle informa-
tion. As a general trend, it can be observed that the improve-
ment of the proposed baseline over the AFE increases as less
oracle information is available. This is coherent with fact that
the AFE was designed for good segmentation conditions, de-
spite having a frame-dropping mechanism, and lower SNRs as
those found on the average microphone. A particularly strong
difference is the use of the GMM used by the baseline in the
case in which no oracles are available.

Effect of single-channel sub-modules: Table 3 details the im-
provements attained with each speech enhancement sub-module
under O3 conditions. Note that the microphone closest to
speaker is used rather than DS for comparability with the AFE.
The largest improvement is obtained estimating noise from the
background frames provided by the AED and a MMSE-LSA es-
timator. This error can be brought down by using IMCRA dur-
ing speech events to improve the estimate of background noise.
The combination of IMCRA with IUP also had an important
effect in performance.

5. Discussion and Conclusions
We have presented a complete toolkit for the evaluation of
multi-channel processing techniques in a distributed micro-
phone scenario. Simulation of reverberant and noisy condi-
tions, although unable to precisely represent all the complex-
ities of a real acoustic scene, allows to define a sufficiently
realistic setup for evaluating and comparing a variety of com-
ponents and algorithms for distant-talking speech recognition.
The results obtained in this multi-channel corpus demonstrate
the challenges presented by distributed microphones in multi-
room environments. They particularly show that considerable
difficulties arise when ASR is extended to a multi-microphone,
multi-command scenario. To address the arising issues, the suc-
cessful design of a specific pipeline of tools integrating VAD or
microphone processing among others, combined with classical
approaches to robust ASR like enhancement or adaptation, can
become challenging.

The proposed package is being constantly updated to foster
research collaboration and algorithms sharing in a realistic and
integrated experimental framework and to provide a state-of-
the-art baseline as a reference for future evaluation campaigns.
The porting to other recognition engines, e.g., Kaldi, is envi-
sioned as well. The extension to real datasets may also con-
tribute to identify the impact of phenomena not currently mod-
eled in our simulations, namely directivity and motion of a real
speaker, rapidly changing acoustic conditions and spontaneous speech.

<table>
<thead>
<tr>
<th>Oracle Conditions</th>
<th>WER (Baseline)</th>
<th>WER (AFE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>74.01</td>
<td>29.09</td>
</tr>
<tr>
<td>O2</td>
<td>46.83</td>
<td>37.88</td>
</tr>
<tr>
<td>O3</td>
<td>33.49</td>
<td>21.34</td>
</tr>
</tbody>
</table>

Table 2: WER comparison of the AFE and the proposed base-
line for all oracle conditions on dev1.

In some cases, e.g., Microphone-Network Processing, no
specific method was developed for the baseline and the DIRHA-
GRID oracle information was used instead. The different levels of
oracle information available are detailed in Table 2. The O1
oracle implied knowledge of loose oracle boundaries for each
speech event within the 1-minute utterance. These boundaries
left a small context similar to, e.g., the AURORA4 or REVERB
tasks. Tighter boundaries are also available but were discarded
as they are less realistic and interfere with adaptation and the
design of some algorithms, i.e., the AFE. The O2 oracle implied
knowledge of the room where the activity took place, whereas O3
implied also knowledge of the exact position of the speaker.

Adaptation and overfitting: As shown by the experimental re-
results and the provided description, the DIRHA-GRID corpus
brings a new level of complexity to robust ASR tasks. One of
the most complicated aspects of the DIRHA-GRID is the dif-
ficulty to reduce the mismatch between the trained models and
the testing environment, which comprises a multi-room, multi-
command environment with a complex processing pipeline.
The results here shown, employ the training and adaptation rou-
tine provided with the DIRHA-GRID. As it can be seen in Ta-
ble 4, despite maintaining the same trend, the WERs are much
higher for the test sets, which is clear evidence of overfitting to
the development data. During the baseline preparation, various
alternative training and adaptation schemes were tested but the
one initially provided was kept. It should be noted however that
more complex adaptation schemes, a.s e.g., reverberant training
using LA6 data and simpler MLLR adaptation, lead to similar
results overall but with a higher WER in the development set.

<table>
<thead>
<tr>
<th>Oracle Conditions</th>
<th>WER (Baseline)</th>
<th>WER (AFE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>41.01</td>
<td>41.87</td>
</tr>
<tr>
<td>O2</td>
<td>13.52</td>
<td>13.52</td>
</tr>
<tr>
<td>O3</td>
<td>37.22</td>
<td>38.53</td>
</tr>
</tbody>
</table>

Table 4: WER on test1 and test2 for the proposed methods for
oracle condition O3.

Adaptation and overfitting: As shown by the experimental re-
results and the provided description, the DIRHA-GRID corpus
brings a new level of complexity to robust ASR tasks. One of
the most complicated aspects of the DIRHA-GRID is the dif-
ficulty to reduce the mismatch between the trained models and
the testing environment, which comprises a multi-room, multi-
command environment with a complex processing pipeline.
The results here shown, employ the training and adaptation rou-
tine provided with the DIRHA-GRID. As it can be seen in Ta-
ble 4, despite maintaining the same trend, the WERs are much
higher for the test sets, which is clear evidence of overfitting to
the development data. During the baseline preparation, various
alternative training and adaptation schemes were tested but the
one initially provided was kept. It should be noted however that
more complex adaptation schemes, a.s e.g., reverberant training
using LA6 data and simpler MLLR adaptation, lead to similar
results overall but with a higher WER in the development set.

5. Discussion and Conclusions
We have presented a complete toolkit for the evaluation of
multi-channel processing techniques in a distributed micro-
phone scenario. Simulation of reverberant and noisy condi-
tions, although unable to precisely represent all the complex-
ities of a real acoustic scene, allows to define a sufficiently
realistic setup for evaluating and comparing a variety of com-
ponents and algorithms for distant-talking speech recognition.
The results obtained in this multi-channel corpus demonstrate
the challenges presented by distributed microphones in multi-
room environments. They particularly show that considerable
difficulties arise when ASR is extended to a multi-microphone,
multi-command scenario. To address the arising issues, the suc-
cessful design of a specific pipeline of tools integrating VAD or
microphone processing among others, combined with classical
approaches to robust ASR like enhancement or adaptation, can
become challenging.

The proposed package is being constantly updated to foster
research collaboration and algorithms sharing in a realistic and
integrated experimental framework and to provide a state-of-
the-art baseline as a reference for future evaluation campaigns.
The porting to other recognition engines, e.g., Kaldi, is envi-
sioned as well. The extension to real datasets may also con-
tribute to identify the impact of phenomena not currently mod-
eled in our simulations, namely directivity and motion of a real
speaker, rapidly changing acoustic conditions and spontaneous speech.
6. References


