

The DIRHA simulated corpus

L. Cristoforetti¹, M. Ravanelli¹, M. Omologo¹, A. Sosi¹,
A. Abad², M. Hagmüller³, P. Maragos⁴

¹Fondazione Bruno Kessler (FBK), Via Sommarive 18, 38123 Povo (TN), Italy
cristofo,mravanelli,omologo,alesosi@fbk.eu

²INESC-ID/IST, R. Alves Redol 9, 1000-029 Lisbon, Portugal
alberto.abad@l2f.inesc-id.pt

³Graz University of Technology, Inffeldgasse 16c, A-8010 Graz, Austria
hagmueller@tugraz.at

⁴Athena Research and Innovation Center, 15125 Maroussi, Greece
maragos@cs.ntua.gr

Abstract

This paper describes a multi-microphone multi-language acoustic corpus being developed under the EC project Distant-speech Interaction for Robust Home Applications (DIRHA). The corpus is composed of several sequences obtained by convolution of dry acoustic events with more than 9000 impulse responses measured in a real apartment equipped with 40 microphones. The acoustic events include in-domain sentences of different typologies uttered by native speakers in four different languages and non-speech events representing typical domestic noises. To increase the realism of the resulting corpus, background noises were recorded in the real home environment and then added to the generated sequences. The purpose of this work is to describe the simulation procedure and the data sets that were created and used to derive the corpus. The corpus contains signals of different characteristics making it suitable for various multi-microphone signal processing and distant speech recognition tasks.

Keywords: data collection, speech contamination, impulse responses

1. Introduction

Natural spontaneous speech interaction with distant microphones is an important step towards the development of easy-to-use voice interfaces in an increasing number of possible applications. With this type of interaction, except for rather controlled real scenarios, present Automatic Speech Recognition (ASR) systems still exhibit lack of robustness and flexibility and, as a result, an unacceptable variability in their performance (M. Wölfel and J. McDonough, 2009). In fact, past activities on this topic showed that state-of-the-art speech recognition systems are inadequate for an interaction at distance from microphones due to several reasons, including the environmental noise, the reverberation effects introduced by the environment and, related to it, the position and head orientation of the speaker.

During the last decade, these topics were investigated under different projects as, for instance, CHIL (Waibel and Stiefelhagen, 2009) and DICIT (Omologo, 2010). In the former case, microphone arrays were distributed in space in order to detect, localize and classify acoustic events, as well as to transcribe speech, typical tasks for contexts as lectures and seminars. In the latter case, the application scenario was spoken dialogue with a TV and related devices, in a typical living-room environment. In both cases, voice interaction was addressed introducing some constraints as the monitoring of a single room, or the possible position of the subject (in front of a TV in the case of DICIT).

The DIRHA project¹, which started in January 2012, addresses the challenge of distant-speech recognition and understanding in a home environment, with a very realistic and complex application scenario as reference, which is

characterized by a fully flexible interaction in any room and position in space. Exploiting a microphone network distributed over the different rooms of an apartment, the targeted system should react properly when a command is given by the user. Moreover, the system is “always-listening”, and an important challenge is to develop a solution that reduces false alarms due to misinterpretation of normal conversations and other generic sounds which do not carry any relevant message to the system.

Related to the above-mentioned functionalities and features, the main fields on which research is conducted are: multichannel acoustic processing, distant speech recognition and understanding, speaker identification/verification, and spoken dialogue management. The final target is to integrate a prototype in real automated homes for an on field evaluation by real users. At this moment, four languages (i.e., Italian, Greek, Portuguese, and Austrian German) are considered, while the English language will be addressed for comparison purposes during the third and last year of the project.

In speech recognition and understanding, in particular in the case of interaction at distance from the microphones, a crucial step regards a proper selection of data and corpora suitable to train and test the various speech processing, enhancement, and recognition algorithms. This follows from the fact that both the size of the corpus and its matching with real usage conditions determine the accuracy of the resulting statistical models used in the recognition step, and eventually the overall system performance. On the other hand, collecting and transcribing appropriate data sets to cover any possible application contexts would become a prohibitive, time-consuming and expensive task. In a domestic context, in particular, due to the large variabil-

¹<http://dirha.fbk.eu>

ities that can be introduced when deploying such systems in different houses, this issue becomes even more critical than in any other traditional ASR application. There is not the possibility to develop corpora that are large and representative enough to cover any real noisy and reverberant conditions in a domestic context under which the above-mentioned technologies might be deployed.

An alternative approach, that is suggested in this case, is to apply multi-condition training and contamination methods (Matassoni et al., 2002), which means that some speech material is derived from existing clean speech corpora and simulation methods. Many past works (see for instance (Huang et al., 2008)) based on the use of contaminated speech showed the convenience and effectiveness of this approach to reduce the mismatch between training and testing conditions, and in case eventually apply adaptation methods to better track the actual conditions under which the interaction with the real end-user runs.

In order to support the research foreseen in the DIRHA project, a set of experimental tasks has been defined. For this purpose, both simulated and real corpora were created. Real data include a collection of distant-speech material as well as acoustic measurements in home environments. Simulated data were produced thanks to a technique that reconstructs, in a very realistic manner, multi-microphone front-end observations of typical scenes occurring in a domestic environment. For this purpose, an automated apartment² is available in Trento, Italy, which represents the site where DIRHA prototypes will initially be developed and tested. Apart from the installed microphones, the apartment features windows, blinds, doors, lights, and heating system controlled by a central unit.

In the following of this paper, the design of the simulated corpus is outlined giving a detailed description of its contents and introducing the basic corpora adopted to generate the simulations. Moreover, the paper aims to describe the tool that was used to generate the simulated multi-channel sequences.

2. The simulated corpus

The DIRHA SimCorpus is a multi-microphone and multi-language database containing simulated acoustic sequences derived from the above mentioned microphone-equipped apartment.

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. Each sequence consists of real background noise with superimposed various localized acoustic events. Acoustic events occur randomly (and rather uniformly) in time and in space (within predefined positions) with various dynamics. The acoustic wave propagation from the sound source to each single microphone is simulated by convoluting the clean signals with the respective impulse response (IR). The IR data collection is described in Section 2.1.

Acoustic events are divided into two main categories, i.e., speech and non-speech. Speech events include different

types of sentences (i.e., phonetically-rich sentences, read commands, spontaneous speech and commands, keywords) uttered by different speakers in the four languages. This speech corpus is described in Section 2.2. Non-speech events have been selected from a collection of high-quality sounds typically occurring within a home environment (e.g., radio, TV, appliances, knocking, ringing, creaking, etc.) and are described in Section 2.3.

For cross-language comparison purposes, each simulated acoustic sequence has been replicated in the four languages while preserving the same background noises and non-speech sources. Gender and timing of the active speakers have been preserved across the different languages, in order to further ensure homogeneity.

Data were generated by means of a multi-microphone simulation framework developed at FBK, as described in Section 2.4.

2.1. The multi-microphone IR corpus

A basic ingredient to generate the simulated corpus is the impulse response measurement, which has been carried out in the reference apartment.

The flat, depicted in Figure 1, comprises five different rooms equipped with a network of 40 omnidirectional microphones (Shure MX391/O). The experimental set-up was based on both distributed microphone networks (pairs or triplets of sensors on the walls) and more compact microphone arrays (on the ceiling of the living-room and the kitchen). In each pair, the sensors are at a distance of 30 cm, while triplets are based on microphones at 15 cm of distance. The arrays are composed of six microphones, five of which have been placed on a circumference of radius 30 cm, while one sensor has been arranged on the center of the circle. One of these arrays is shown in Figure 2.

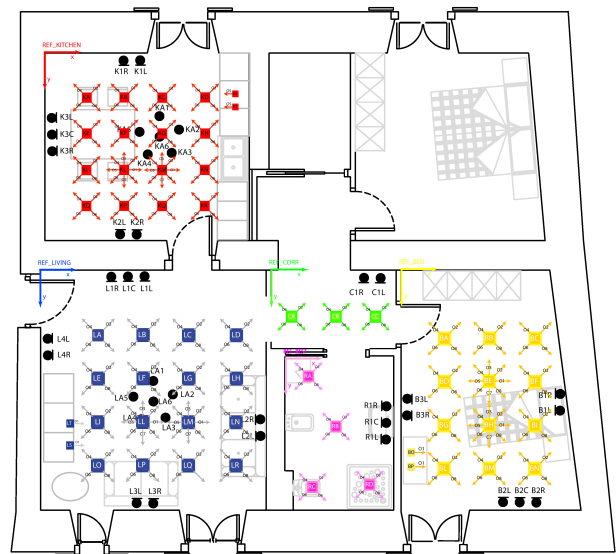


Figure 1: An outline of the microphone set-up adopted for the DIRHA SimCorpus. Black dots represent microphones, while boxes and arrows represent available positions and orientations of the loudspeaker used to measure the IRs.

²Property of Istituto Trentino per l'Edilizia Abitativa – ITEA S.p.A.



Figure 2: A six-element microphone array installed on the ceiling.

The IRs measurement process explored several different positions and orientations of the loudspeaker over the various rooms of the apartment, in order to generate simulated data with a satisfactory level of richness in terms of spatial variability. As shown in Table 1, more than 9000 sample-synchronized IRs have been measured. Cross-room impulse responses are also included in order to simulate sources in other rooms.

The huge number of measured IRs derives from the multiplication of the various orientations and positions by the total number of microphones. As an example, in the bathroom 16 positions/orientations (see Figure 1) were recorded by 40 microphones, leading to a number of 640 different IRs.

The IRs measurements were based on a professional studio monitor (Genelec 8030A) able to excite the target environment with long sequences of Exponential Sine Sweep (ESS) signals (Farina, 2000). As pointed out in (Ravanelli et al., 2012), ESS method ensures IRs measurements with a high Signal to Noise Ratio (SNR) and a remarkable robustness against harmonic distortions.

The Time of Flight (TOF) information, crucial to applications such as acoustic event localization, beam-forming, and multi-microphone signal processing in general, has been preserved by means of six sample-synchronized multi-channel audio cards (RME Octamic II). The measured IRs are at 48kHz sampling frequency with 24-bit accuracy.

2.2. The multi-language clean speech corpus

A clean-speech data set has been created to derive the simulated corpus outlined in the previous section. For this purpose, very high quality close-talking speech signals, leading to at least 40dB of SNR for each sentence, have been recorded at each partner site, in order to avoid any possible artifacts otherwise introduced by convolution with impulse responses. For each language, either 20 or 30 speakers were recorded. These speakers, aged between 25 and 50, have an equal gender distribution. The subjects were sitting in front of a display in the recording room. The recording process

Room	Installed mics	Available positions	IRs	T_{60} (s)
Living-room	15	18	2960	0.74
Kitchen	13	18	2960	0.83
Bedroom	7	14	2160	0.68
Bathroom	3	4	640	0.75
Corridor	2	3	480	0.60
Total	40	57	9200	

Table 1: Some details on the microphone network adopted for the DIRHA SimCorpus. The last column reports the estimated reverberation time T_{60} , which indicates that the room acoustic characteristics are quite challenging for distant-talking ASR studies.

was controlled by an operator outside the room, in order to show on the display each sentence to read.

Recordings were performed using professional microphones (Neumann U89i and TLM 103, AKG C 414B-UL, and Studio Projects T3 Dual Triode) with a pop filter. Particular care has been taken to avoid any interferences, both acoustical and electrical.

The subjects read different typologies of data, in order to augment the recorded material also beyond the purposes of the DIRHA project. The Austrian German data collection has been recorded as a part of a bigger corpus (GRASS), described in (Schuppler et al., 2014).

Each speaker read the following material:

- phonetically-rich sentences (i.e., sentences designed to have a large phone coverage and phonetic context);
- read commands (i.e., typical commands to operate in-house devices);
- spontaneous commands (i.e., typical commands to operate in-house devices, after visualizing some domestic pictures on the display);
- keywords (i.e., open-sesame keywords);
- conversational speech (i.e., free speech on a given topic).

Sentences containing errors in speech pronunciation have been deleted from the clean data set. All the sentences (except for the conversational speech sentences) were annotated in text files, following a standard previously adopted for TIMIT (Garofolo et al., 1993), APASCI (Angelini et al., 1994), and other similar corpora.

2.3. The non-speech events and background noises corpus

In order to generate realistic acoustic sequences, high-quality recordings of typical home noisy events were needed. In the context of these simulations, we chose more than 300 different sounds (appliances, knocks, ringing, creaking, and other similar noises) from the Freesound³ and

³<http://www.freesound.org/>

Apple Logic Pro database. A selection of copyright-free radio shows, music and movies were used to simulate radio and television sounds.

To further increase the realism of the acoustic sequences, a set of 16 common background noise sequences (i.e., shower, washing machine, oven, vacuum, etc.) were directly recorded, in a multi-channel fashion, in the reference apartment and then properly added to the convoluted signals. The resulting multi-channel noise sequences are in this way extremely realistic since they are directly recorded from the target acoustic environment.

2.4. The simulation process

The contamination process adopted to generate the simulations is depicted in Figure 3. A set of dry acoustic events (speech or typical home noises) are selected from the available clean corpora. For each source, a random position in space is chosen, and then a convolution between the clean signal (with gain randomly chosen within a specified range) and the proper set of multi-microphone IRs is performed to account for the room acoustics. To increase the realism of the acoustic sequences, real background noise sequences of different dynamics (noise gain) have also been added. The speech sentences might be overlapped in time with any other acoustic source (i.e., a phone ring, a door bell, music, and all the other possible localized noises). The overlap between speech and non-speech acoustic sources is allowed both in the same and between different rooms, while the overlap between speech acoustic sources occurs mostly in different rooms.

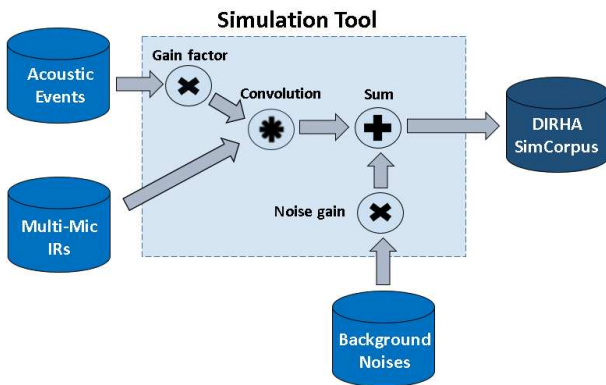


Figure 3: Basic scheme of the simulation process adopted for the generation of the DIRHA SimCorpus.

The simulation software is written in MATLAB and it is easily reconfigurable to account for new acquisition setups or different acoustic event distributions.

2.5. Corpus features

Each sequence of each language includes a set of speech events (i.e., a keyword followed by a read command, a spontaneous command, a phonetically-rich sentence, a segment of conversational speech) and a variable number of localized non-speech sources.

For each language, the simulated corpus is divided into two or three chunks (DEV, TEST1, TEST2) containing 75

acoustic sequences each with 10 different speakers involved for each data set. Table 2 shows the exact composition of each data set, while some details on the total size of the corpus and an overall statistic in terms of SNR averaged over all the sequences are respectively reported in Table 3 and Table 4.

Language	DEV (seq)	TEST1 (seq)	TEST2 (seq)	TOT (seq)
Italian	75	75	75	225
German	75	75	-	150
Greek	75	75	-	150
Portuguese	75	75	-	150
TOT	300	300	75	675

Table 2: Available sequences for the DIRHA SimCorpus. Since each sequence lasts 60 seconds, the number of available sequences corresponds to the number of minutes. TEST2 is available only for the Italian language, since 30 speakers were recorded only for that language.

Language	DEV (GB)	TEST1 (GB)	TEST2 (GB)	TOT (GB)
Italian	61	58	60	179
German	62	60	-	122
Greek	62	60	-	122
Portuguese	60	58	-	118
TOT	245	236	60	541

Table 3: Total size of the DIRHA SimCorpus in GB. The disk space occupation accounts for all the 675 acoustic sequences by considering all the 40 channels (48 kHz, 16 bits) in mixed, separated and clean conditions.

Language	DEV (dB)	TEST1 (dB)	TEST2 (dB)
Italian	16.0±11.7	17.2±11.6	16.7±11.6
German	15.2±11.7	16.7±11.8	-
Greek	15.6±11.8	16.6±11.6	-
Portuguese	15.1±11.6	16.7±11.5	-

Table 4: Mean SNR of the acoustic sequences and standard deviation. For each speech source, the SNR is computed and averaged over all the microphone signals related to the same room. In the corpus the standard deviation is rather high, since there is a significant variability in the environmental noise conditions, as a typical domestic scenario.

The simulated corpus contains both mixed sources (i.e., the multi-microphone signals containing all the acoustic sources mixed-up) and separated sources (i.e., the single source separated by the other acoustic sources). The clean (or dry) signals of each acoustic source are also included in the corpus. Figure 4 highlights the differences among the three categories of available signals.

An XML annotation file, available for each microphone, describes each sequence specifying each simulated condition in detail. An excerpt of an annotation file is reported in Figure 5, showing an acoustic event occurring in the

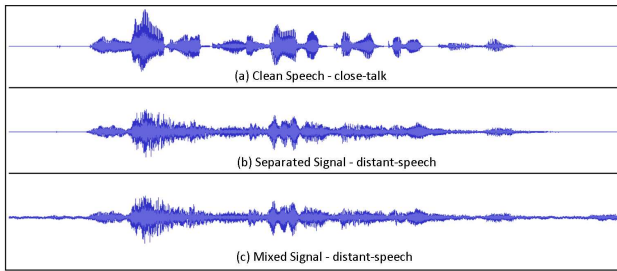


Figure 4: Each source of each acoustic sequence is available in three different scenarios of increasing complexity. In (a) is reported a read command uttered in a clean and anechoic environment. The same sentence has been available (for all the sensors of the microphone network) in the reverberated version (b) and in the most challenging scenario (c) in which both noises and reverberation degrade the quality of the signal.

living-room. In the example, the name of the source under description (`<name>`), the begin and end samples, and the adopted impulse response, as well as position of the source, are reported. Additional fields describe the reverberation time (`<T60>`), the SNR, the gender and ID of the speaker, and the transcription of the uttered sentence. The documentation is completed by additional information that describe the overall geometric configuration, the segmentation boundaries related to each acoustic event, and a pictorial representation of the geometry of the acoustic sequence.

```
<SOURCE>
  <name>sp_ph_rich</name>
  <begin_sample>770838</begin_sample>
  <end_sample>984848</end_sample>
  <IR>LIVINGROOM/LG/O2/LA1.mat</IR>
  <pos>xs=295 ys=188 zs=150 REF=REF_LIVINGROOM</pos>
  <T60>0.75s</T60>
  <SNR>23.61dB</SNR>
  <gender>Female</gender>
  <SPK_ID>spk6</SPK_ID>
  <label>txt>
    770838 984848 nel giardino c' e' una statua muliebre
  </label>txt>
</SOURCE>
```

Figure 5: Excerpt of a XML annotation file describing a phonetically-rich sentence uttered in the living-room by a female speaker.

3. Possible applications

Several experiments conducted in DIRHA have shown the usefulness of the corpus for testing of source localization, acoustic event detection, speech/non-speech discrimination techniques. Thanks to the presence of clean, mixed, and separated signals, the corpus is also suitable for multi-microphone acoustic echo cancellation. Furthermore, such a database is also adequate for multi-microphone speech enhancement and for distant speech recognition.

Finally, the split of the corpus in three different chunks including different speakers, helps the user perform parameter tuning on the development set while the other sets can be used for test purposes.

In the context of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014), a special session dedicated to “Speech detection and speaker localization in domestic environments” (Brutti et al., 2014) has been proposed. In particular, a portion of the DIRHA SimCorpus has been delivered (together with some real data sequences extracted from Wizard of Oz experiments) to researchers working in the field of multi-microphone signal processing in order to assess their algorithms in this domestic scenario.

4. Conclusions

The DIRHA SimCorpus is a huge collection of simulated acoustic sequences, suitable for various signal processing and speech recognition tasks.

We plan to extend the corpus by including more languages such as English and French. Furthermore, we will explore some other microphone configurations involving, for instance, arrays with a different geometric configuration and, possibly, a higher number of microphones.

The corpus will be made available to the research community under possible future initiatives, such as international challenges and special sessions. At the moment an excerpt of six simulations extracted from the SimCorpus is already available for downloading through the project website.

5. Acknowledgments

This work was partially funded by the European Union, Seventh Framework Programme for research, technological development and demonstration, grant agreement no. FP7-288121, under DIRHA. We would like to thank some colleagues that helped us in the data collections, in particular A. Brutti and P. Svaizer (FBK), F. Batista and A. Costa (INESC-ID), B. Schuppler, H. Pessentheiner and J. Cordovilla (TU Graz), Z. Skordilis, I. Rodomagoulakis, P. Giannoulis, A. Katsamanis and G. Potamianos (Athena R.C. IAMU).

6. References

- B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. 1994. Speaker independent continuous speech recognition using an acoustic-phonetic Italian corpus. In *Proceedings of ICSLP 1994*.
- A. Brutti, M. Ravanelli, P. Svaizer, and M. Omologo. 2014. A speech event detection/localization task for multiroom environments. In *Proceedings of HSCMA 2014*.
- A. Farina. 2000. Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. In *Proceedings of 108th AES Convention*.
- J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. 1993. DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM.
- J. Huang, M. Epstein, and M. Matassoni. 2008. Effective Acoustic Adaptation for A Distant-talking Interactive TV System. In *Proceedings of Interspeech 2008*.
- M. Wölfel and J. McDonough. 2009. *Distant Speech Recognition*. Wiley.
- M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer. 2002. HMM Training with Contaminated Speech Material for Distant-Talking Speech Recognition. *Computer Speech and Language*.
- M. Omologo. 2010. A prototype of distant-talking interface for control of interactive TV. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*.
- M. Ravanelli, A. Sosi, P. Svaizer, and M. Omologo. 2012. Impulse response estimation for robust speech recognition in a reverberant environment. In *Proceedings of EU-SIPCO 2012*.
- B. Schuppler, M. Hagmüller, J. A. Morales Cordovilla, and H. Pessentheiner. 2014. GRASS: the Graz corpus of Read And Spontaneous Speech. In *Proceedings of LREC 2014*.
- A. Waibel and R. Stiefelhagen. 2009. *Computers in the Human Interaction Loop*. Springer.