THE DIRHA-ENGLISH CORPUS: an overview on the dataset with the related tools and recipes

Mirco Ravanelli, Maurizio Omologo

Fondazione Bruno Kessler (FBK), 38123 Povo, Trento, Italy

{mravanelli,omologo} @ fbk.eu

Distant Speech Recognition (DSR) may represent, in the future, a valid alternative to standard close-talking ASR [1]. In effect, although the latter approach usually leads to better performance, it is easy to predict that users will prefer to relax the constraint of handling or wearing any microphone-equipped device to access speech recognition services. There are indeed various real-life situations where a distant-talking (hands-free) interaction would be more natural convenient and attractive. An emerging field is, for instance, represented by speech based domestic control, which has been largely explored under the European Project DIRHA [1].

Despite the growing interest towards DSR, current technologies still exhibit a lack of robustness and flexibility since the presence of non-stationary noises and acoustic reverberation typically make DSR particularly challenging [2].

In order to foster the scientific progress in this field and to establish a common framework across researchers, the availability of realistic corpora may play a crucial role for the development and assessment of DSR solutions. Along this line, the considerable success of some international challenges such as REVERB [3], CHIME [4] and ASpIRE, has contributed to the wide diffusion of common corpora, tasks and baselines. Nevertheless, we feel that other complementary corpora and tasks can be fruitful to the research community, in order to explore some scenarios which are not fully explored by previous data-sets.

The DIRHA-English corpus [5], along with the other data-sets developed under DIRHA [6,7,8,9], gives the chance to assess novel techniques considering a very large number of microphone channels, including both microphone networks and microphone arrays distributed in a domestic environment. Moreover, the corpus allows researchers to compare their techniques on different types of microphones, ranging from high-quality condenser microphones to cheap digital MEMS microphones. The overall data-set consists of different lists of commands and keywords, as well as wsj5k, wsj20k, phonetically-rich and conversational sentences, offering the possibility to explore tasks of different complexity. Some portions of the corpus are going to be publicly distributed [2]. Together with the data-set, kaldi recipes and tools will be made available to reproduce baseline results.

[1] The research presented here has been partially funded by the European Union 7th Framework Programme (FP7/2007-2013) under grant agreement no. 288121 DIRHA (for more details, please see http:// dirha.fbk.eu).

[2] The modalities for downloading the corpus will be made under *http://dirha.fbk.eu*

The purpose of this demo is to provide information that is complementary to what presented in [5] giving a deepen insight into the DIRHA-English corpora, for instance showing in detail how the simulated and real sequences were generated. In particular, we will start describing our methodology and tools for multi-microphone impulse response computation based on the Exponential Sine Sweep (ESS) technique [10]. We will then show our approach and codes for the generation of the multi-microphone sequences, showing how high-quality close-talking recordings, impulse responses, and recorded noisy sequences can effectively be combined to obtain very realistic simulations. After a listening session of the generated data, we will also address corpus annotation, with a particular emphasis on the tools and strategies adopted to obtain a reliable lexicon and precise phone-level alignments of the generated sequences. The last part of the demo will be devoted to the detailed description of the available tools and kaldi recipes to run baseline DSR experiments.

REFERENCES

[1] M. Wolfel and J. McDonough, Distant Speech Recognition, Wiley, 2009.

[2] E. Hansler and G. Schmidt, Speech and Audio Processing in Adverse Environments, Springer, 2008.

[3] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, “The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech,” in Proc. of WASPAA 2013, 2013, pp. 1–4.

[4] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines,” in Proc. of ASRU 2015, 2015.

[5] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, M. Omologo “The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments,” in Proc. of ASRU 2015, 2015.

[6] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmueller, P. Maragos, "The DIRHA simulated corpus", in Proc. of LREC 2014, 2014.

[7] M. Matassoni, R. Astudillo, A. Katsamanis, M. Ravanelli, “The DIRHA-GRID corpus: baseline and tools for multi-room distant speech recognition using distributed microphones”, in Proc. of INTERSPEECH 2014, 2014.

[8] E. Zwyssig, M. Ravanelli, P. Svaizer, M. Omologo, "A multi-channel corpus for distant-speech interaction in presence of known Interferences", in Proc. of ICASSP 2015, 2015.

[9] A. Brutti, M. Ravanelli, P. Svaizer, M. Omologo, "A speech event detection and localization task for multiroom environments", in Proc. of HSCMA 2014, 2014.

[10] A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” in Proc. of the 108th AES Convention, 2000, pp. 18–22.