
Proceedings Paper:
Mokaram, S. and Moore, R.K. orcid.org/0000-0003-0065-3311 (2017) The Sheffield Search and Rescue corpus. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5-9 March 2017, New Orleans, USA. IEEE , pp. 5840-5844. ISBN 9781509041176

https://doi.org/10.1109/ICASSP.2017.7953276

THE SHEFFIELD SEARCH AND RESCUE CORPUS

Saeid Mokaram, Roger K. Moore

Speech and Hearing Research Group, Department of Computer Science, University of Sheffield, UK
{s.mokaram, r.k.moore}@sheffield.ac.uk

ABSTRACT

As part of an ongoing research into extracting mission-critical information from Search and Rescue speech communications, a corpus of unscripted, goal-oriented, two-party spoken conversations has been designed and collected. The *Sheffield Search and Rescue* (SSAR) corpus comprises about 12 hours of data from 96 conversations by 24 native speakers of British English with a southern accent. Each conversation is about a collaborative task of exploring and estimating a simulated indoor environment. The task has carefully been designed to have a quantitative measure for the amount of exchanged information about the discourse subject. SSAR includes different layers of annotations which should be of interest to researchers in a wide range of human/human conversation understanding as well as automatic speech recognition. It also provides an amount of data for analysis of multiple parallel conversations around a single subject. The SSAR corpus is available for research purposes.

Index Terms— conversational speech corpus, goal-oriented conversation, spoken language understanding, automatic speech recognition

1. INTRODUCTION

Recent years have witnessed significant improvements in the technology of automatic speech recognition. This has led to new interests in both academic and commercial worlds into the processing of natural spoken conversations and ultimately assimilation of their information content. By far the most common place to find such interest is in tracking meetings, analysing customer service calls and extracting their valuable information (such as topics discussed, decisions made, customer satisfaction) for management purposes. However recently, some attentions have been drawn into the role of processing speech communication channels in more critical and challenging application domains such as emergency services (fire, ambulance, etc.) and crisis intervention centres to provide valuable situational knowledge for better decision-makings [1].

This work was supported by the University of Sheffield Cross-Cutting Directors of Research and Innovation Network (CCDRI), Search and Rescue 2020 project.

Despite the existence of language resource agencies such as LDC[1] and ELRA[2], limited natural human/human spoken data is available for research purposes due to issues such as privacy, copyright or signal quality. For Spoken Language Understanding (SLU) tasks, the situation is even worse. The construction of understanding systems using statistical approaches requires suitable annotated data. For measuring the performance of the information extraction systems, it would be ideal if each conversation contains a quantitative amount of information about the discourse subject. In addition, due to the diverse nature of understanding tasks, datasets often needed to be tailor-made to their specific needs.

Since 1990, when the term SLU was coined by ATIS project [2], a variety of speech corpora has been collected. Whilst the majority of these corpora were designed for the more constrained task of human/machine interactions, some notable attempts such as Switchboard [3] and Fisher [4] provide a good amount of two-party human/human conversational speech data. They have been extensively used in their original targeted research areas of speech and speaker recognition rather than speech understanding or information extraction. Call-Home [5] and Call-Friend [6] were collected in response to the need for more natural and multilingual/accented conversational speech data. In the context of crisis response, the PRONTO corpus [7] (in German) was collected from voice communications in exercise missions by the Dortmund Fire Department, Germany. The collection is specifically designed to study the impact of terrestrial trunked radio codecs on keyword extraction and speech recognition. Other recent collections, – AMI corpus [8] and DARPA-funded CALO [9] – were designed to study extensions of human/human conversations such as meetings, lectures, and broadcasts. In contrast to these corpora in which the dialogues are about general random topics, the Maptask [10], TRAINS [11] and Monroe [12] corpora are collections of task-oriented dialogues. The Monroe corpus, in particular, consists of a relatively rich dialogue domain because of its larger and more complex task of disaster handling compared to the simple tasks of giving directions on a paper map in the Maptask and transportation planning in TRAINS. These collaborative tasks were designed to study natural human dia-

[1] Linguistic Data Consortium (LDC): www.ldc.upenn.edu

[2] European Language Resources Association (ELRA): www.elra.info

logue behaviours. However, they are less concerned about the information content of dialogues about the discourse subject.

This paper presents a new corpus of unscripted, goal-oriented, two-party spoken conversations which has been designed, recorded and transcribed as part of an ongoing research into extracting mission-critical information from speech communication channels within the Search and Rescue (SAR) context [13, 14]. The *Sheffield Search and Rescue* (SSAR) was made based on an abstract communication model between First Responders (FRs) and Task Leaders (TLs) during search process in a crisis response training scenario. Each conversation is concerned with a cooperative task of exploring a simulated indoor environment by FR and estimating a topological map of the environment by TL via asking FR about their observations. While the dialogues are spontaneous and participants were free to talk about the simulated environment, an implicit constraint is applied to these conversations by the task and the environment structure as the discourse subject. The environment structure has carefully been designed in order to have a quantitative measure for the amount of exchanged information in each conversation about the discourse subject. The level of map estimation accuracy by a TL can be expressed as the information content of the conversation.

The SSAR corpus is available for research purposes. The full instruction on reproducing the recording setup together with simulated environments are also released which allows future attempts in the expansion of this speech corpus.

2. CONVERSATION TASK DESIGN

2.1. Conversation scenario

Speech communications in the SAR context is a good example of human/human conversation. It is a complex communication scenario with the principal intention of exchanging information between rescue agents and synchronizing their knowledge about an incident scene. Fig. 1 illustrates an abstract model of the communications between FRs and TLs using a pictographic visual language introduced in [15]. In this model, the FRs' goal is to explore the environment and report their observations back to the control hub to update the TL's knowledge about the incident scene. This abstract model was used to design the underlying task for the SSAR conversations.

The SSAR task involves two participants in the roles of an FR and a TL. To simulate a remote conversation, they are located in separate quiet rooms. Wearing headsets, the TL is able to hear FRs reports and talk back for asking or confirming any required information. The FR is the main speaker in this task and speaks most of the times reporting to the TL about the incident scene, their observations and actions. Given pen and paper and just relying on these explanations, the TL is asked to make an estimation of the structure of simulated environment by drawing nodes to represent rooms/locations and links between them to show how they

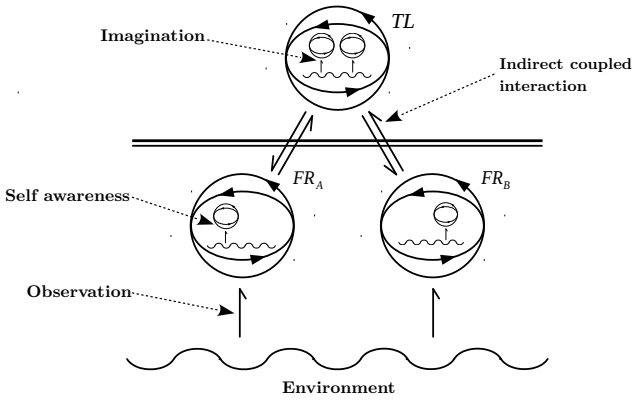

Fig. 1. *Pictograph illustration of the abstract communication model within the SAR context.*

might be connected to each other. The TL is also asked to annotate each node by writing down some of the key features (e.g. room type and its condition, or key objects and their characteristics) about each location in a way that each node can be identified from the others. The final goal is to have an estimated topological map of the incident scene.

2.2. Simulated environment and maps design

Inspired by the simulation training systems (e.g. FLAMESIM [16]) which are being used by some fire departments to practice their communication performance and decision making, a simulated indoor environment was designed and built in Unity 3D game engine [17]. The designed simulation system is similar to a first-person-shooter 3D game in which a participant can explore the simulated environment by moving an avatar around using arrow keys on the keyboard. Fig. 2 shows a user-view of the simulated environment.

Fig. 2. *A user-view of the designed simulation system.*

In the SSAR task, the conversations are centered around transferring enough information about the environment from FR to the TL in order to describe its general structure. This indicates that the design of the environment map is of particular importance because, more complex is the structure, more information is required to be transferred over speech channel during a successful conversation. In other words, the struc-

tural complexity of an environment map can affect the information content of a conversation.

An approach for studying the complexity and information stored in a structure is to describe it as a graph. A generic structural model has been used to make the environment maps clear and not too complicated to describe. In this model, each structure comprises numbers of square rooms which can be connected to each other by doors. These structure of connected rooms can easily be symbolized by an undirected graph which its nodes represent the rooms and links between them indicate the doorways. While all the rooms have an identical square shape, different objects and arrangements inside them give a unique identity to each.

The graph entropy, which is commonly used as the structural information content and the complexity of a graph [18, 19], is used to design four different map settings with a range of complexity. The topological structure of these four map settings are shown in Fig. 3. Each map setting consists of fix number of 8 rooms. Some maps have multiple rooms with the same type; for example Map_2 has two bedrooms; however, different objects and arrangements inside them gives a unique identity to each. In total 13 different types of indoor locations ($RoomTypes$), such as *kitchen*, *bedroom* or *computer lab*, were simulated in all four map settings.

Various types of ambient noises (e.g. fire noise, washing machine noise, boiler room noise, etc.) were also simulated which the FR can hear in *stereo* form to provide a realistic experience. The TL can also hear these background noises in the FR's environment with a -20dB level difference and in *mono* in order to simulate a natural telephone conversation.

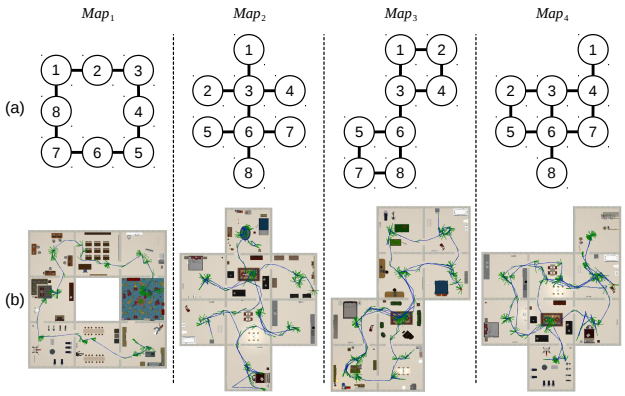

Fig. 3. *(a) The topological structure of four different map settings ($Map_{1\text{-}4}$) which were explored by each participant; (b) corresponding top-view image of each map which are overlaid with the motion trajectory of a participant and her viewing directions (small arrows) at each time.*

3. SSAR CORPUS RECORDING

Recordings were performed in two separate quiet rooms for avoiding external acoustic disturbances and crosstalk between the two speakers' voice. Fig. 4 illustrates a schematic of this

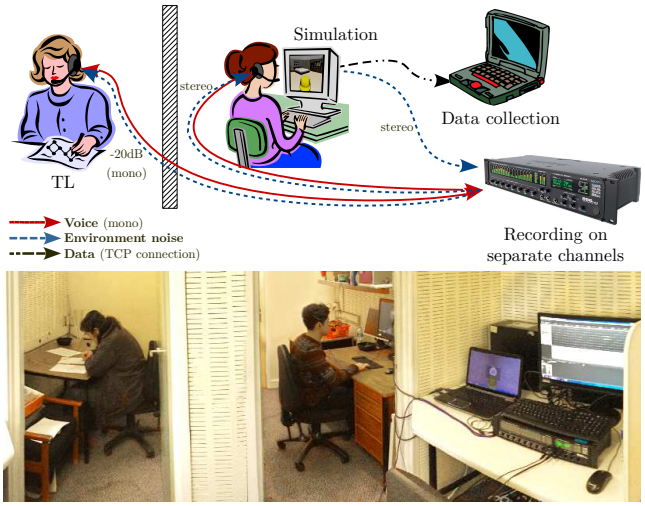

Fig. 4. *top: the recording scenario, bottom: the recording set-up in two separate quiet rooms.*

setup (top) and a photo was taken from two participants while starting a recording session (bottom). The participants performed the experiment behind the closed doors by communicating with each other through the simulated remote communication system. A *MOTU-896Mk3* [20] audio interface/mixer was used to provide the simulated communication system by mixing the participants voices and the background environment noise with their appropriate loudness levels for each speaker. This interface system, together with *Audacity* [21] software, was used for A/D conversion and recording the speakers' voice and the simulated environment noise on four separate channels; one channel for each participant and two for environment noise (i.e. *stereo*). Other information about participants' motion trajectories, actions, and list of objects in their field of view in the environment were logged in a computer readable text file.

Each recording was started by the participant in the role of an FR by pressing a *connect* button in the simulation GUI. A maximum time for each map was estimated based on some practice recordings during the process of the conversation task design. Maximum tasks duration were set as 6, 7, 8 and 8 minutes for Map_1, Map_2, Map_3 and Map_4 respectively. In order to motivate the participants to explore and explain the maps accurately, they were offered an additional cash reward to their volunteering fee for estimating each map correctly. The majority of the participants explored the entire area of each map and there were just about 12.5% who could not manage to visit all the rooms in the limited time. In all experiments, the structure of the explored area of the environment was correctly estimated by the TLs. Fig. 5 presents a hand drawing example of the Map_4 estimated by a participant. Correct estimation of the visited areas confirms that the amount of exchanged information through voice channel is sufficient for a human subject to estimate the structure the visited parts of the environment.

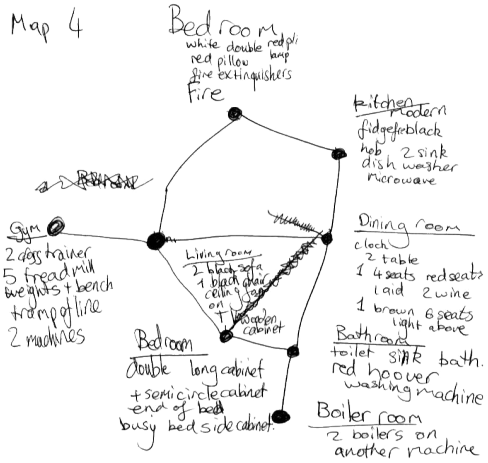

Fig. 5. *A hand drawing example of the Map_4 estimated by a participant (TL).*

4. TRANSCRIPTION AND ANNOTATION

The segmentation and transcription are generated automatically in a first round based on Automatic Speech Recognition (ASR) transcriptions of clean speech data. The ASR system used for the first round transcription was accessed through webASR [22]. Its outputs were then reformatted to XML files compatible with Transcriber [23] for an accurate manual transcription. Then the segmentations and transcriptions were revised by a trained native English speaker in a format compatible with the rules in the AMI corpus [24]. Fig. 6 presents some sections of a conversation between an FR and a TL as an example of the conversations and their transcripts.

Each transcription file has been included with the recording meta-data comprising subjects' gender, age and accent region together with information about the map setting, starting room and conversation duration. More detailed information about each recording has also been provided in a separate *TASK-INFO* text file. The corpus perplexity against a standard *Switchboard* 3-gram Language Model (LM) was 173. About 11% of utterances contain at least a token indicating aspiration, cough/throat clearing, laugh or other prominent vocal noises.

5. CORPUS DESCRIPTION

SSAR is a medium size multi-speaker corpus with 96 two-party goal-oriented spoken conversations lasting from 6 to 8 minutes duration (averaging $\sim$7.25 minutes) each. A total of 24 native British English speakers (66.6% Male) with a southern accent (self-reported) participated in the recording. All the participants were recruited as paid volunteers through the Sheffield-student-volunteers system. The corpus totals about 12 hours of speech data and $\sim 80K$ words of manual transcription with $\sim 16K$ vocabulary size, $\sim 11K$ utterances and $\sim 1K$ dialogue turns. Each speaker's clean speech and the environment noise are available on separate channels. This enables the researchers to have more control over the background noise by altering the noise level or even removing or replacing it with other noises.

[. . .]
FR er i'm going through one of the other doors that I haven't been through yet
FR er this is a bedroom
TL okay
FR there is a bed a double bed
FR there is a bedside table | with what's either a mirror or a picture
[. . .]
FR okay i'm going | there's no more doors going off from this room
TL okay
FR so i'm going back into the dining room with the tables and i'm going through the only other door I haven't been through yet
TL yep
FR er this looks like a_ | wash erm | a toilet or washing room
FR er there are no doors going off from this one
FR there is a bath | with a curtain
[. . .]
FR I think that's everything
FR er | on your map is there any rooms I haven't explored yet
TL erm yeh | from the library there's two rooms
TL if you go from the dining room to the living room
FR okay | yep
TL and | from there | oh sorry from the living room there is two rooms
FR okay I see | okay there is a another bedroom it's a child's b_ with a child's bedroom
FR there is a_desk with a lamp
[. . .]

Fig. 6. *Some sections of a conversation between a FR and a TL as an example of the conversations and their transcripts in the SSAR.*

Aligned with these recordings other information about the participants' locations, actions and objects in their field of view in the environment are available on computer readable log-files. This information can be used as a form of conceptual annotation for the conversations. Multiple layers of annotations in this corpus would be of interest to researchers in a wide range of human/human conversation understanding tasks as well as ASR. The SSAR corpus also provides an adequate amount of data for analysis of multiple parallel conversations around a single subject. The current version does not include dialogue act tagging annotation. The spoken conversations have many of the characteristics of spontaneous spoken language such as disfluencies, false starts, and colloquial pronunciations.

6. CONCLUSION

New interests are emerging in both academic and commercial worlds into the processing of natural spoken conversations and automatic extraction of their information content. This has led to a demand for new speech corpora of unscripted, goal-oriented, meaningful spoken conversations. Corpora of human-human conversations are required that each dialogue is guaranteed to contain a quantitative amount of information explaining a particular discourse subject. We are publishing the Sheffield Search and Rescue corpus in response to this need. The SSAR corpus is available for research purposes. The full instruction on reproducing the recording setup together with simulated environments are released to encourage future attempts in the expansion of this speech corpus.

7. REFERENCES

[1] D. V. Kalashnikov, D. Hakkani-Tür, G. Tür, and N. Venkatasubramanian, "Speech-Based Situational Awareness for Crisis Response," in *EMWS DHS Workshop*, 2009.

[2] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The ATIS Spoken Language Systems Pilot Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 96–101, 1990.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, San Francisco, CA, mar 1992, pp. 517–520.

[4] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, vol. 4, Lisbon, Portugal, may 2004, pp. 69–71.

[5] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech LDC97S42," DVD, Philadelphia, 1997.

[6] A. Canavan and G. Zipperlen, "CALLFRIEND American English-Non-Southern Dialect LDC96S46," Web Download, Philadelphia, 1996.

[7] D. Stein and B. Usabaev, "Automatic Speech Recognition on Firefighter TETRA broadcast," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[8] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI Meeting Corpus: A Pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 28–39.

[9] V. Pallotta, J. Niekrasz, and M. Purver, "Collaborative and argumentative models of meeting discussions," in *Proceeding of CMNA-05 international workshop on Computational Models of Natural Arguments*, 2005.

[10] A. H. Anderson, M. Bader, E. G. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, and J. Miller, "The HCRC map task corpus," *Language and Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[11] J. Allen and P. Heeman, "TRAINS Spoken Dialog Corpus LDC95S25," CD, Philadelphia, 1995.

[12] A. J. Stent, "The Monroe Corpus," University of Rochester, Rochester, NY, USA, Tech. Rep., 2001.

[13] S. Mokaram and R. K. Moore, "Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario," in *Proceedings of Interspeech*. ISCA, 2015, pp. 2734–2738.

[14] S. Mokaram and R. K. Moore, "Speech-based topological map estimation in a simulated search and rescue environment," in *NIPS workshop on Machine Learning for Spoken Language Understanding and Interaction*, Montreal, 2015.

[15] R. K. Moore, "Introducing a pictographic language for envisioning a rich variety of enactive systems with different degrees of complexity," *International Journal of Advanced Robotic Systems*, vol. 13, no. 2, p. 74, 2016.

[16] FLAME-SIM, "FLAME-SIM: Fire department training simulation software," 2016. [Online]. Available: http://www.flame-sim.com

[17] Unity, "Unity (Personal): a 3D game engine development platform." 2016. [Online]. Available: https://unity3d.com/

[18] M. Dehmer and F. Emmert-Streib, "Structural information content of networks: Graph entropy based on local vertex functionals," *Computational Biology and Chemistry*, vol. 32, no. 2, pp. 131–138, 2008.

[19] A. Mowshowitz and M. Dehmer, "Entropy and the complexity of graphs revisited," *Entropy*, vol. 14, no. 3, pp. 559–570, 2012.

[20] MOTU-896Mk3, "An audio interface/mixer," 2016. [Online]. Available: http://www.motu.com/products/motuaudio/896mk3

[21] Audacity, "Audacity: A free, open source software for recording and editing sounds," 2016. [Online]. Available: http://www.audacityteam.org/

[22] T. Hain, A. El Hannani, S. N. Wrigley, and V. Wan, "Automatic speech recognition for scientific purposes - WebASR," in *Proceedings of Interspeech*. Brisbane, Australia: ISCA, 2008, pp. 504–507.

[23] C. B. Liberman, E. Geoffrois, Z. Wu, and Mark, "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech," in *First International Conference on Language Resources and Evaluation (LREC)*, 1998, pp. 1373–1376. [Online]. Available: http://trans.sourceforge.net

[24] J. Moore, M. Kronenthal, and S. Ashby, "AMI transcription," 2016. [Online]. Available: http://groups.inf.ed.ac.uk/ami/corpus/transcription.shtml