The Sheffield Wargames Corpus

Charles Fox[1], Yulan Liu[1], Erich Zwyssig[2,3], Thomas Hain[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, S1 4DP, UK
[2]Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, UK
[3]EADS IW, Appleton Tower, Edinburgh, EH8 9LE, UK

Abstract

Recognition of speech in natural environments is a challenging task, even more so if this involves conversations between several speakers. Work on meeting recognition has addressed some of the significant challenges, mostly targeting formal, business style meetings where people are mostly in a static position in a room. Only limited data is available that contains high quality near and far field data from real interactions between participants. In this paper we present a new corpus for research on speech recognition, speaker tracking and diarisation, based on recordings of native speakers of English playing a table-top wargame. The Sheffield Wargames Corpus comprises 7 hours of data from 10 recording sessions, obtained from 96 microphones, 3 video cameras and, most importantly, 3D location data provided by a sensor tracking system. The corpus represents a unique resource, that provides for the first time location tracks (1.3Hz) of speakers that are constantly moving and talking. The corpus is available for research purposes, and includes annotated development and evaluation test sets. Baseline results for close-talking and far field sets are included in this paper.

1. Introduction

Automatic Speech Recognition of clean speech has matured to provide practical commercial applications such as word processor dictation and phone voice menu selection. However recognising speech in natural environments such as business meetings and casual conversations is still an unsolved problem. A useful application of meeting recognition in conjunction with language processing methods would be the automated production of minutes for meetings. Meetings have several challenging properties for ASR: speakers are typically seated at several places around a table, microphones can generally only be placed at a few fixed locations away from the seating locations so recognition must be by farfield methods; speakers may speak simultaneously during heated discussions; and – most seriously for current beamforming methods – speakers may sometimes move around while speaking, such as walking to a whiteboard to present ideas to the group. The latter breaks the beamforming and source-separation assumptions of stationary sources used in many systems, and the development of algorithms to handle this case is an active research area. However there has been a lack of realistic data to train and test such algorithms, for two reasons. First, real business meetings have been unwilling to make public their contents as they often concern sensitive information. Meetings have been simulated by actors such as in the AMI corpus [3] but it is unknown how they compare to real meetings. Second, within available recorded meetings, the amount of data containing moving speakers is usually small, as the bulk of most meetings has the speakers seated and stationary. It is difficult to obtain large amounts of realistic data for moving speakers in meetings, despite the need for recognition of this type of data, which often contains the most valuable content of the meetings (a speaker standing to present is often more informative than one sitting during a general discussion).

To aid research in this area we have collected, and are making public as part of this publication, a 7 hour recording of the natural speech in a realistic environment where the speakers are almost constantly moving and often talking over one another. To avoid the above problems of meeting recordings we examined several surrogate scenarios with the requirement that the speech be from a real task (not reading text, not artificial tasks or by actors) which encourages movement, and by native English speakers. After examining several types of teamwork and game scenarios, we found that tabletop wargames provide an ideal surrogate. This paper describes the Sheffield Wargames Corpus (SWC) collected from 7 hours of tabletop gaming as a surrogate for mobile speakers during business meetings. We have pre-processed the corpus and defined evaluation sets and development sets for future users, and we present the initial baseline result using standard farfield ASR methods.

1.1. Related work

The AMI corpus [3] provides 100 hours of meeting recordings, a mixture of real and acted meetings, and includes basic speaker movement data for some parts. The non-acted meetings are for an artificial design task which does not yield the same level of interactivity as the wargames task; there is also no music in the background and particuants do not move around as much as when playing a tabletop game. AMI particpants often do not know each other and interact formally, in SWC, in several cases players have played together for up to 10 years and speak very casually and naturally, for example completing each others' sentences. The ICSI Meeting Corpus [8] provides natural meeting data from up to 12 participants wearing lapel microphones, and from four desk mounted microphones. The NIST RT09 challenge included meeting speech recognition sets from several sources (NIST,CMU etc.) but equivalently meetings were mostly not equivalent to the AMI style scenarios, athough some included board game playing.

The COSINE corpus [11] provides 145 hours (26.7 transcribed speech) of noisy indoor and output conversational fluent (though some non-native) English speech by groups of 2-7 paid volunteers equipped with head, shoulder, throat and chest-array microphones. Subjects were requested to have natural conversations with eath other and a list of topics provided. Unlike COSINE, the Sheffield Wargames Corpus does not require natural speech for its own sake, which may still be somewhat artificial in a recording envionment, but records natural speech of players in a real meeting-like task where the speech is essential for the task. Our subjects did not appear to be affected by the presence

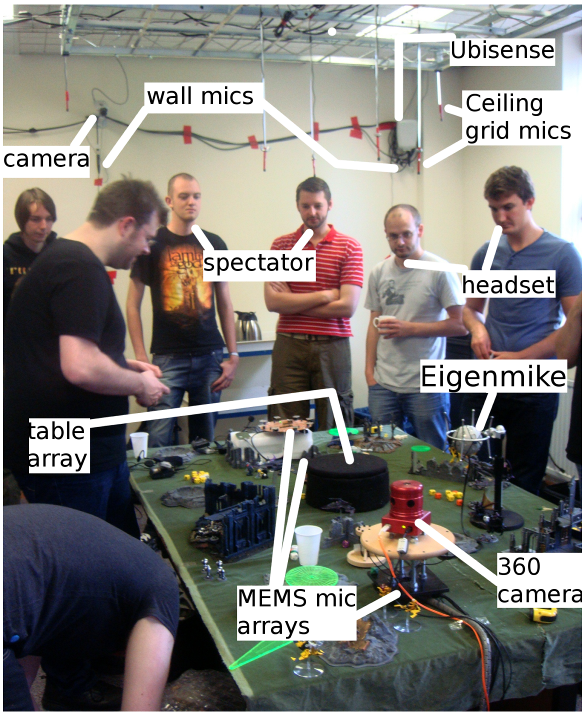


Figure 1: Wargame players during the recording. Mics can be seen hanging from the ceiling grid, on the back wall, on headsets, and in arrays on the table incorporated into the game scenery.

of recording equipment in any way, and many commented that they felt normal wearing their headsets. Many players were also online gamers who wear similar headsets online. COSINE does not include location data.

Corpora such as [10] and [9] exist containing natural speech in the specific noisy environment inside cars. Natural conversations over telephones are available in [6] and other corpora. The CHIME corpus [4] provides read speech in from a stationary location with natural speech and domestic background noise. The CAVA [2] corpus provides binaural recordings of moving speakers. None of these corpora however include location data or moving speakers relative to fixed microphones.

2. Corpus Description

2.1. Wargame Recording

The corpus recorded the speech of players of a tabletop game named Warhammer 40,000 [12]. In each game, four players (in two teams of two players) wear headsets to which the Ubisense[1] tracking tags adhere. Each game lasts for around 1.5 hours and comprises several (typically four) turns. Each turn consists of three formal phases in which different activities are carried out (moving, shooting and fighting with miniature soldiers). Team members must talk to one another; measure distances with a tape-measure throughout their actions to decide what to do; roll dice; and talk with the opposing team to agree on the effects of their actions upon them and vice versa. Fig. 1 shows a typical moment from a game. To maximise players' movement, the room was cleared of seating so players stand and walk around.

Game	Ses.	Dur.	#Utt.	Notes
1	1	38:00	441	–
	2	44:59	323	–
2	3	33:54	352	–
	4	34:49	352	–
	5	38:00	172	–
3	6	47:79	407	–
	7	32:10	237	–
	8	42:59	295	Dev set 3 pizza party
4	9	35:50	301	Background rock music.
	10	-	-	Excluded from corpus.
	11	41:49	404	Background light music.

Table 1: Wargame recording summary, showing game number, duration and number of utterances in each session.

They were also asked to customise the rules of their games to create as much motion as possible (e.g. banning fixed artillery but allowing fast vehicles which require players to physically move them long distances around the table.) At various times, additional spectators enter and leave the room without any head microphones, but they may also speak.

The whole corpus consists of 4 games in 10 release sessions (Table 1). The first 8 sessions of the first 3 games are mainly recorded without background noise, while the remaining sessions include background music (rock music and light music). In part of Session 8, Game 3, a 'party' was held by inviting all the day's players and some recording assistants to eat and chat while a game was in progress. A total of nine male native British English speakers participated in the recording. Some of the sessions include between-game speech such as setting up, packing away, and chatting. We originally recorded 11 hours of data but have cropped to 7 hours to remove recording problems such as incorrectly-worn head microphones and equipment failures.

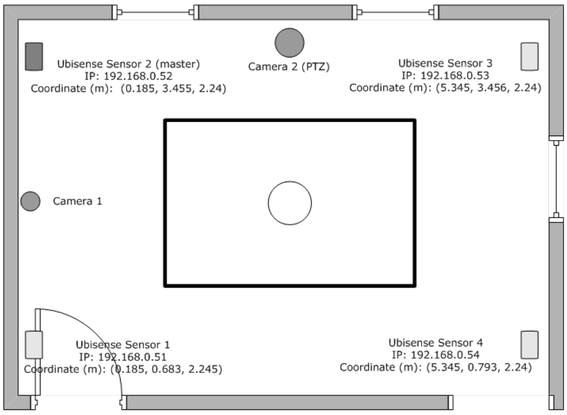
There are 96 audio channels in total. Twenty-four sample-synchronous microphones recordings are composed of 4 groups: 4 headsets on the player, 4 microphones adhered to the wall, one circular microphone array of 8 on the table, one microphone array hanging on the grid to the ceiling (figure 2) in the shape of two overlapping squares. All distant microphones are hyper-cardioid AKG C417/III vocal condenser microphones, while the headsets are all Sennheiser ew100 wireless headsets of cardioid directivity. These 24 channels are recorded sample-synchronous using an all-Linux setup. 48kHz, 16 bit A/D conversion is by MOTO 8Pre's, linked by firewire 400 and FFADO drivers to JACK middleware on an Ubuntu Studio desktop PC, streaming audio data direct to hard disc. Full details of the sample-synchronous recording setup can be found in [5].

Additionally, but not sample-synchronised to the above channels, we made further recordings using an omnidirectional 32-channel Eigenmike© array, five 8-channel microphone arrays using analogue and digital MEMS microphones and the PointGrey Ladybug2 camera. A detailed description of this setup is available from the homepage of the 2012_MMA corpus [2].

There are two cameras hanging overhead (figure 3) and a 360 degree panoramic camera on the table recording simultaneously with the microphones. Four Ubisense 3D location sensors are installed at corners of the ceiling to collect signals from

[1] Ubisense: http://www.ubisense.net/en/about-us

[2] http://www.cstr.inf.ed.ac.uk/research/#corpora



Ubisense Sensor 2 (master)
IP: 192.168.0.52
Coordinate (m): (0.185, 3.455, 2.24)
Camera 2 (PTZ)
Ubisense Sensor 3
IP: 192.168.0.53
Coordinate (m): (5.345, 3.456, 2.24)
Camera 1
Ubisense Sensor 1
IP: 192.168.0.51
Coordinate (m): (0.185, 0.683, 2.245)
Ubisense Sensor 4
IP: 192.168.0.54
Coordinate (m): (5.345, 0.793, 2.24)

Figure 2: Audio recording configuration

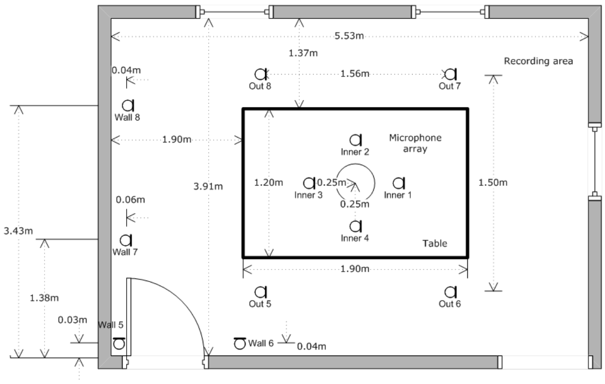


Figure 3: Video and location tracking system configuration

Ubisense tags attached to players' headsets, i.e. to track their head locations.

The Eigenmike and MEMS mics, video and Ubisense were synchronised manually to the 24 sample-synchronous audio channels by aligning a Ubisense-tracked clapper-board's clap at the start of each session.

2.2. Ubisense Location Tracking

In the public release data we provide both the raw data from the Ubisense location tracking system and a smoothed version. The manufacturer's stated precision of the raw recordings is 15cm and the effective location update rate is about 1.3Hz. To compensate for location noise, simple exponential smoothing is performed by selecting a exponential factor to compromise between smoothing error and the smoothness with the *Least Normalised Distance* (LND) method, where the *surge*, (third order derivative) of the smoothed data is used as a metric of smoothness. The curve between the smoothing error and the surge has a cross point with each coordinate axis respectively, then both axis are scaled and normalised so that the cross point is unit 1. The corresponding exponential smoothing factor to the point with least distance from the origin on the curve is selected to smooth the raw location data (Figure 4).

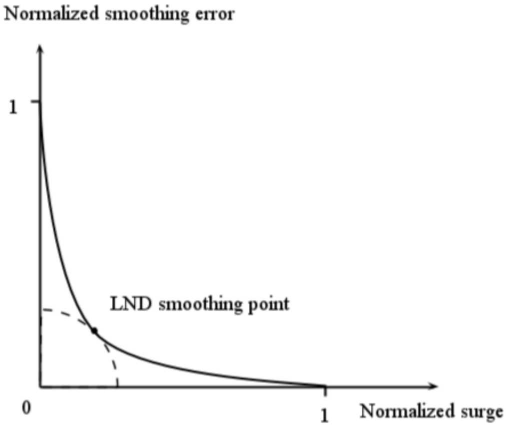


Figure 4: Exponential smoothing factor determination with LND method

2.3. Pre-processing and Set Selection

In each session, 8 and 7 minutes of data are selected for the evaluation and development sets respectively. The evaluation set is selected from the middle of a session when players are talking most seriously and when there are only few artefacts in the recording channel. Development sets are selected before and after the evaluation sets similarly by avoiding heavy laughs, breaths or recording artefacts as much as possible.

Segmentation, annotation and transcription were performed with *XTrans*[3]. For the evaluation sets, manual segmentation, annotation and transcription are generated in a first round of annotation based on headset recordings. Then the transcription is revised by native English speakers and the format is corrected to be compatible with rules as used in the AMI corpus[4] and published in NIST STM format. Currently the development sets have been manually segmented and annotated based on headset recording. We also publish meta-data comprising subjects' English nativeness and accent region, gender, age and height.

85% of utterances overlap with at least one other (1.11 hours of 1.35hours) in the transcribed evaluation set. The perplexity of the evaluation set against a standard AMI 3-gram language model [7] was 167, with an OOV rate of 1.3% (=203 of 15430 words). The reason for the relatively high perplexity and low OOV seems to be that many entities in the game have known English words as names but they are used in unusual ways, for example soldiers called 'space wolves'. There are 3255 utterances in the test set, 26% contain a LAUGH or BREATH token (7% LAUGH, 18% BREATH).

2.4. Segment Tagging

Based on raw data from the Ubisense tracking system, the evaluation data was grouped according to the location and the moving speed. Four spatial criteria are used and all are based on segmental average of location, distance and moving speed. This is to enable standard comparisons of algorithms such as beamformers whose performance may depend on location. *Spatial Splitting Circular* (SSC) is based on the segmental average distance from the speaker active in that headset channel segment to

[3] LDC tools XTrans: `http://www.ldc.upenn.edu/tools/XTrans`

[4] AMI transcription: `http://corpus.amiproject.org/documentations/transcription`

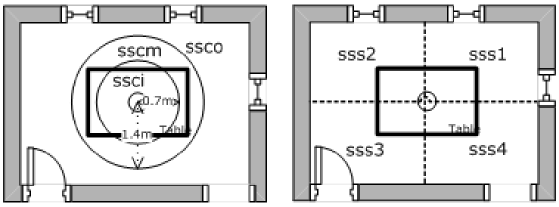


Figure 5: Video and location tracking system configuration

total	ssci	sscm	ssco
2548	281	1660	607
–	11.03 %	65.16 %	23.82 %

Table 2: Statistics of segment tagging: SSC

the centre of the circular microphone array on the table in horizontal plane (figure 5, left). If the distance is less than 0.7m, it's tagged as "ssci" where "i" stands for "inner circle". If the distance is larger than 1.4m, it's tagged as "ssco" where "o" stands for "outside circle". Otherwise the segment is tagged as "sscm" where "m" stands for "middle circle". Table 2 shows the statistics of the splitting based on the first criterion. Note that the segments with only breath and laugh are excluded here. *Spatial Splitting Square* (SSS) is based on the segmental average location in the room. The room is split into four square parts equally in horizontal plane. As shown in figure 5 (right), the four subareas are named respectively "sss1", "sss2", "sss3" and "sss4", in the same order of mathematical coordinate quartiles. Table 3 gives the statistic details in this case. *Spatial Splitting Height* (SSH) is based on the segmental average vertical height. According to the height of the speakers and the relative position of tags to human head, if the segmental average height is smaller than 1.476m, the segment is tagged with "sshl" where "l" stands for "low" since there is a high possibility the person leans down in that segment. Otherwise it's tagged with "sshn" where "n" stands for "normal". Table 4 gives the statistic details in this case. *SPeeD splitting* (SPD) is based on the segmental average moving speed in horizontal plane. If it's smaller than 0.2m/s, the segment is tagged with "spds" where "s" stands for "slow". Otherwise the segment is tagged with "spdn" where "n" stands for "normal". Table 5 gives the statistic details in this case.

In addition to spatial tags, we also provide tags indicating the session number (which contains information about background music) and whether transcribed utterances overlap.

3. Baseline speech recognition results

For the head microphones and a single distant microphone (SDM, grid channel 0), recognition was performed on individual channels. Acoustic models and language models trained on the AMI corpus [3] and from the AMI RT'09 meeting transcription system [7] were used in the experiments. For the latter both individual head microphone (IHM) data and multiple distance microphone (MDM) models are available, used in the first passes of the RT'09 systems for recognition of IHM and MDM conditions. For the sample-synchronous table-mics-only (table) and table+grid+all (tgw) sets we applied the Beamformit beamformer (BF) [1] with default parameters and no noise reduction pre-processing, to extract a continuous single channel. We used a variety of standard decoders and input features [13] to report a range of results.

total	sss1	sss2	sss3	sss4
2548	512	673	754	609
–	20.09 %	26.41 %	29.59 %	23.90 %

Table 3: Statistics of segment tagging: SSS

total	sshl	sshn
2548	848	1700
–	33.28 %	66.72 %

Table 4: Statistics of segment tagging: SSH

total	spds	spdn
2548	1325	1223
–	52.00 %	48.00 %

Table 5: Statistics of segment tagging: SPD

Scoring was performed with NIST tools although for the far field conditions prior knowledge about the word to channel association was assumed. The results are shown in table 6. We found no significant differences in performance between the music and non-music sessions, and between overlappiing and non-overlapping speech, for the head-mic baselines. For the SDM case only we performed a run on a subset of the data tagged as non-overlapping speech (noOverlap).

4. Discussion

Speech recognition research is moving to more natural situations, but the community has lacked a natural, distant, native English speech corpus of highly mobile speakers with location data and we are publishing the Sheffield Wargames Corpus to fill this need. Our baselines show this is a highly challenging ASR task – harder than previous location corpora using actors – even with state of the art adapted CMVN/HLDA/CMLLR systems with well trained acoustic and language models achieving only 65.3 WER from the head mics and 85.8 from a basic default beamformer. These baselines will provide a challenge both for our and others' research on natural distant speech and the publication of 96 audio channels and location data should aid development of more powerful natural speech technology.

5. Acknowledgements

Thanks to Sheffield University Wargames Society for their great support, planning and cooperation; Ghada Al Harbi, Davide Marino, Sarah Al Shareef, Raymond Ng and Madina Hasan for recording help. Funded by EPSRC Natural Speech Technology Programme Grant EP/I031022/1 and EADS.

Data	Models	P	WER	S	D	I
IHM	AMI	1	73.7	49.0	18.3	6.4
IHM	AMI	2	70.7	43.6	23.5	3.6
IHM	RT'09 IHM	2	65.3	41.5	19.3	4.5
SDM-o1	RT09 MDM	2	87.3	44.2	40.7	2.4
table-o1	RT09 MDM	2	86.8	40.6	43.8	2.4
gtw-o1	RT09 MDM	2	85.7	39.8	43.0	2.9

Table 6: Baseline word error rate results on the SWC evaluation test set. P = number of passes.

6. References

[1] Xavier Anguera, Chuck Wooters, and Javier Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech and Language Processing*, (7):2011–2023, 2007.

[2] Elise Arnaud, Heidi Christensen, Yan-Chen Lu, Jon Barker, Vasil Khalidov, Miles Hansard, Bertrand Holveck, Hervé Mathieu, Ramya Narasimha, Elise Taillant, Florence Forbes, and Radu P. Horaud. The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In *ACM/IEEE International Conference on Multimodal Interfaces (ICMI'08)*, October 2008.

[3] J.C. 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 *Machine Learning for Multimodal Interaction, Second International Workshop, 11-13 July 2005, Edinburgh, UK. pp. 28-39. Lecture Notes in Computer Science 3869. Springer Verlag*, 2006.

[4] Heidi Christensen, Jon Barker, Ning Ma, and Phil Green. The CHiME corpus: a resource and a challenge for computational hearing in multisource environments. In *Proc. Interspeech*, 2010.

[5] C. Fox, H. Christensen, and T. Hain. Linux audio for multi-speaker natural speech technology. In *Proc. Linux Audio Conference*, 2012.

[6] J. Godfrey, E. Holliman, and J. McDaniel. SWITCHBOARD: telephone speech corpus for research and development. In *Proc. ICASSP*, 1992.

[7] Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Asmaa el Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan. The AMIDA 2009 Meeting Transcription System. In *Interspeech'10*, pages 358–361, 2010.

[8] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The ICSI meeting corpus. In *Proc. ICASSP*, 2003.

[9] N. Kawaguchi, S. Matsubara, H. Iwa, S. Kajita, K. Takeda, F. Itakura, and Y. Inagak. Construction of speech corpus in moving car environment. In *Proc. ICASSP*, 2000.

[10] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang. Avicar: audio-visual speech corpus in a car environment. In *Proc. ICSLP*, 2004.

[11] Alex Stupakov, Evan Hanusa, Jeff Bilmes, and Dieter Fox. Cosine - a corpus of multi-party conversational speech in noisy environments. In *Proc. ICASSP*, 2009.

[12] Games Workshop. *Warhammer 40,000 Rulebook, 5th Ed.* Games Workshop Ltd. ISBN 978-1841548753, 2008.

[13] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland. *The HTK Book, version 3.4.* Cambridge University Engineering Department, Cambridge, UK, 2006.