

# Deriving and Exploiting Situational Information in Speech: Investigations in a Simulated Search and Rescue Scenario

Saeid Mokaram



Department of Computer Science  
The University of Sheffield

*PhD Thesis*  
*submitted for the degree of Doctor of Philosophy*  
*Supervisor: Professor Roger K. Moore*



***Dedicated***

***to***

***my parents***

*for always supporting me with affection and love  
despite the distance.*

***my beloved wife Hamideh***

*for her endless love.*





# Abstract

---

The need for automatic recognition and understanding of speech is emerging in tasks involving the processing of large volumes of natural conversations. In application domains such as Search and Rescue, exploiting automated systems for extracting mission-critical information from speech communications has the potential to make a real difference.

Spoken language understanding has commonly been approached by identifying units of meaning (such as sentences, named entities, and dialogue acts) for providing a basis for further discourse analysis. However, this fine-grained identification of fundamental units of meaning is sensitive to high error rates in the automatic transcription of noisy speech. This thesis demonstrates that topic segmentation and identification techniques can be employed for information extraction from spoken conversations by being robust to such errors.

Two novel topic-based approaches are presented for extracting situational information within the search and rescue context. The first approach shows that identifying the changes in the context and content of first responders' report over time can provide an estimation of their location. The second approach presents a speech-based topological map estimation technique that is inspired, in part, by automatic mapping algorithms commonly used in robotics. The proposed approaches are evaluated on a goal-oriented conversational speech corpus, which has been designed and collected based on an abstract communication model between a first responder and a task leader during a search process. Results have confirmed that a highly imperfect transcription of noisy speech has limited impact on the information extraction performance compared with that obtained on the transcription of clean speech data.

This thesis also shows that speech recognition accuracy can benefit from rescored its initial transcription hypotheses based on the derived high-level location information. A new two-pass speech decoding architecture is presented. In this architecture, the location estimation from a first decoding pass is used to dynamically adapt a general language model which is used for rescored the initial recognition hypotheses. This decoding strategy has resulted in a statistically significant gain in the recognition accuracy of the spoken conversations in high background noise.

It is concluded that the techniques developed in this thesis can be extended to more application domains that deal with large volumes of natural spoken conversations.



# Declaration

---

I hereby declare that I am the sole author of this thesis. The contents of this thesis are my original work and have not been submitted for any other degree or any other university. I have designed and collected the speech dataset described in Chapter 3. Some parts of the work presented in Chapters 3, 4 and 5 have been published in conference proceedings as given below:

1. Saeid Mokaram and Roger K. Moore, “Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario”, in *Interspeech*, 2015. (Oral presentation)
2. Saeid Mokaram and Roger K. Moore, “Speech-Based Topological Map Estimation in a Simulated Search and Rescue Environment”, in *NIPS 2015, Workshop on Machine Learning for Spoken Language Understanding and Interaction*, 2015. (Poster presentation)
3. Saeid Mokaram and Roger K. Moore, “The Sheffield Search and Rescue Corpus”, in the IEEE International Conference on *Acoustics, Speech, and Signal Processing* (ICASSP), 2017. (Poster presentation)
4. Saeid Mokaram, Hamideh Kerdegari, Christina Georgiou, Roger K. Moore, Tony J. Prescott, Tony J. Dodd, “Search and Rescue 2020”, University of Sheffield Engineering Symposium 2013 group project. (First prize in group poster competition).
5. Saeid Mokaram and Roger K. Moore, “High-Level Context for Improving Automatic Recognition of Conversational Speech”, Submitted to *Interspeech*, 2017.



# Acknowledgements

---

First and foremost I would like to express my sincere gratitude to Roger Moore for giving me the opportunity to work with him. As I slowly developed in the role of a researcher, Roger has patiently provided me with his wealth of knowledge and insight on the subject of speech technology. I consider myself extremely lucky to have had his guidance and advice throughout my PhD study.

I gratefully acknowledge the University of Sheffield Cross-Cutting Directors of Research and Innovation Network (CCDRI), Search and Rescue 2020 project, which provided funding for this work.

My wonderful time working towards this thesis could have been very different without the many who have supported me on the way. I would like to thank Jon Barker, Lucia Specia, and Tom Stafford for their support during my PhD. I would like to thank Tony Dodd for introducing me to the ‘Search and Rescue 2020’ network project.

I would like to thank Rosanna Milner for kindly helping me in editing this thesis. I also would like to thank her for helping me in testing my recording set-up and the main recordings.

I am also grateful to my excellent colleagues in the Speech and Hearing research group (SpandH) for bringing a great atmosphere and for their sincere friendship.

Last but not least, I would like to thank my family and particularly my wife Hamideh, without whose love and encouragement the completion of this thesis would not have been possible.

*Saeid Mokaram*

*Sheffield, January 2017*



# Table of Contents

---

<b>List of Figures</b>	<b>xv</b>
<b>List of Tables</b>	<b>xxiii</b>
<b>Abbreviations</b>	<b>xxv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Problem definition . . . . .	5
1.3 Aims and objectives . . . . .	6
1.4 Organization of the thesis . . . . .	7
<b>2 Background</b>	<b>11</b>
2.1 Speech communications in search and rescue . . . . .	12
2.1.1 Speech technology for search and rescue . . . . .	15
2.1.1.1 Situation awareness . . . . .	15
2.1.1.2 Speech-based situation awareness . . . . .	17
2.1.2 Challenges in the automatic processing of voice channels	18
2.2 Automatic speech recognition . . . . .	22
2.2.1 Language model . . . . .	25
2.2.1.1 N-gram model estimation: . . . . .	26
2.2.1.2 Language model adaptation . . . . .	27
2.2.1.3 Model interpolation . . . . .	30
2.2.2 Performance of speech recognition systems . . . . .	31
2.2.3 Multipass decoding . . . . .	34
2.2.3.1 Language model lattice rescoring . . . . .	35

2.3	Understanding speech conversations . . . . .	36
2.3.1	Challenges using speech input . . . . .	37
2.3.2	Topic segmentation . . . . .	39
2.3.2.1	Segmentation techniques . . . . .	40
2.3.2.2	Evaluation metric . . . . .	44
2.3.2.3	Segmentation performance . . . . .	45
2.3.3	Topic identification . . . . .	46
2.3.3.1	Technical approaches . . . . .	50
2.4	Topological mapping . . . . .	59
2.4.1	Topological maps . . . . .	60
2.4.2	Automatic topological mapping . . . . .	61
<b>3</b>	<b>Sheffield Search and Rescue Corpus</b>	<b>65</b>
3.1	Suitable speech communication datasets . . . . .	66
3.2	Conversation task design . . . . .	69
3.2.1	Conversation scenario . . . . .	69
3.2.2	Simulated environment design . . . . .	71
3.3	Corpus recording . . . . .	74
3.4	Transcription and annotation . . . . .	79
3.5	Corpus description . . . . .	79
<b>4</b>	<b>Locational Information Extraction</b>	<b>85</b>
4.1	Speech-based location estimation . . . . .	86
4.1.1	Transition detection . . . . .	88
4.1.2	Location estimation . . . . .	91
4.1.3	The impact of transcription errors . . . . .	92
4.1.3.1	Automatic transcription system . . . . .	93
4.1.3.2	Document classification system . . . . .	94
4.1.3.3	Results . . . . .	96
4.2	Speech-based topological map estimation . . . . .	100
4.2.1	New node detection . . . . .	101
4.2.2	Correspondence estimation . . . . .	102
4.2.3	Experiments and results . . . . .	107



<b>5</b>	<b>Design of a Two-Pass Speech Recognizer</b>	<b>113</b>
5.1	System architecture . . . . .	114
5.2	Experimental set-up . . . . .	116
5.2.1	Baseline automatic speech recognition system . . . . .	116
5.2.2	Experiments . . . . .	117
5.3	Results and discussion . . . . .	120
<b>6</b>	<b>Conclusions</b>	<b>127</b>
6.1	Reviewing the scope of the thesis . . . . .	127
6.2	Answer to research questions . . . . .	131
6.3	Original contributions . . . . .	132
6.4	Future work . . . . .	132
6.5	Conclusion . . . . .	135
	<b>References</b>	<b>137</b>
	<b>Appendix A Examples of TF-IDF</b>	<b>159</b>
	<b>Appendix B Examples of LDA</b>	<b>161</b>
	<b>Appendix C SSAR Corpus Recording Forms</b>	<b>167</b>
	C.1 Information sheet . . . . .	168
	C.2 Personal information form . . . . .	172
	C.3 Consent form . . . . .	173
	<b>Appendix D SSAR Corpus Transcription Guidelines</b>	<b>175</b>
	<b>Appendix E Examples of Manually Estimated Maps</b>	<b>187</b>
	<b>Appendix F ASR Performance at Different SNRs and LMSFs</b>	<b>193</b>
	<b>Appendix G List of Publications and Presentations</b>	<b>197</b>



# List of Figures

---

1.1	Three parts of the ‘ <i>Search and Rescue 2020</i> ’ network project. . .	2
2.1	<i>top: An overview of a typical fire response communication scenario. bottom: the voice and language parameters at each stage of the fire response process (visualized on top) are presented with a focus on characterizing the difficulty level of the ASR task. The triangle shaped bar indicates the ASR task difficulty at each stage. SCBA stands for Self-Contained Breathing Apparatus. . .</i>	13
2.2	<i>A general architecture of a large vocabulary continuous speech recognizer. . . . .</i>	23
2.3	<i>The process of making a context-dependent language model in the rescoring pass. The background language model is typically a large model trained on out-of-domain data and the context-specific language models are typically small but specific models trained on subsets of the training corpus. A context-dependent model can be made dynamically by interpolating these static models based on some information about changes in the speech domain. . . .</i>	32
2.4	<i>An example showing a reference utterance transcript, its ASR hypothesis, and each word error types. C, I, D, and S stand for correct, insertions, deletions and substitutions, respectively. . . .</i>	33
2.5	<i>The general architecture of a two-stage speech decoding. High-level language model or a more sophisticated acoustic model can be used in a second-stage decoding to rescore and re-rank the initial decoding hypotheses. . . . .</i>	35

2.6	<i>TextTiling graph. Vertical lines indicate actual topic boundaries. The graph indicates computed similarity of adjacent windows of text. Peaks indicate coherency, and valleys indicate potential breaks between tiles (reproduced with permission from Hearst and Plaunt (1993)).</i>	41
2.7	<i>a: Dot-plotting of four concatenated Wall Street Journal articles (reproduced with permission from (Reynar, 1994)). b) Utterance similarity plot for a Physics lecture, with vertical lines indicating true segment boundaries (reproduced with permission from (Malioutov and Barzilay, 2006)).</i>	42
2.8	<i>Graphical representation of three primary constraints describing a topic identification task, with example tasks for various combinations of these constraints (Hazen, 2011) (reproduced with permission from Wiley Books).</i>	49
2.9	<i>Block diagram of the four steps typically taken by a speech-based topic identification system during the process for converting an audio document into topic hypotheses (Hazen, 2011) (reproduced with permission from Wiley Books).</i>	50
2.10	<i>Graphical model representation of LDA. The boxes are plates representing replicates. The outer plate represents <math>M</math> documents, while the inner plate represents the <math>N</math> repeated choice of topics (<math>z</math>) and words (<math>w</math>) within a document (Blei et al., 2003) (reproduced with permission from JMLR).</i>	55
2.11	<i>Level of abstraction hierarchy for maps (Boal et al., 2014) (reproduced with permission from Cambridge university press).</i>	60
2.12	<i>The basic steps that are typically taken by automatic topological map making techniques in the field of mobile robotics. The first step is to choose the appropriate technologies to sense the environment while a robot explores an area. The next step is to detect when a new node (landmark) should be added to the map. The final step is to determine whether each added node is a new one, or one that has been visited previously.</i>	62

3.1	<i>Pictograph illustration of the abstract communication model within a search and rescue context. In this model, an individual is represented by a circle and an inner ellipse. The inner ellipse represents its thoughts and understandings. For instance here, the Task Leader (TL) has an understanding (imagination) about First Responders' (FR) status and the environment which they are in. However, each FR has an understanding about themselves and their surrounding environment (self awareness). The double arrows represent the coupled interaction between FRs and the TL that is performed remotely via a voice communication channel. In this model, FR goal is to explore the environment (i.e. incident scene) and report their observations and actions back to the control hub to update the TL knowledge about the incident scene. . . . .</i>	70
3.2	<i>(a) A user-view of the designed simulation system. (b) A top-view of the simulated environment (Map<sub>3</sub>) which is overlaid with the motion trajectory of a participant and their viewing directions (small arrows) at each time. . . . .</i>	72
3.3	<i>(a) The topological structure of four different map settings (Map<sub>1-4</sub>) which were explored by each participant. (b) corresponding top-view image of each map. . . . .</i>	73
3.4	<i>top: the recording scenario, bottom: the recording set-up in two separate quiet rooms. . . . .</i>	75
3.5	<i>An example of motion trajectory information plotted over the environment map, an instance of a participant's field of view and surrounding objects in the simulated environment. . . . .</i>	77
3.6	<i>A hand drawing example of the Map<sub>4</sub> estimated by a participant (task leader). . . . .</i>	78
3.7	<i>Some sections of a conversation between an FR and a TL as an example of the conversations and their transcripts in the SSAR. A '↓' indicates a long (about one second) pause and a '%' token indicates cough/throat clearing. . . . .</i>	80

- 
- 4.1 *Visualisation of self-similarity plot for one example Map<sub>1</sub> conversation transcript in the SSAR corpus. Cosine similarity scores every pair of utterances are presented with a gray levels ranging from white for zero (no similarity), to black for one (highly similar). Red dashed lines show ground-truth transitions between rooms. . . . .* 88
- 4.2 *A typical example of transition estimation on the automatic transcript of a conversation in the SSAR corpus. A sliding window with the size of three was used. The ground-truth and the estimated location transition lines are plotted. The blue line shows the transition class membership probability estimated by the SVM classifier. . . . .* 90
- 4.3 *A typical example of the location identification on the automatic transcript of a conversations in the SSAR corpus. This shows the SVM class membership probability distribution for 13 RoomType classes estimated for each segment. In this example a participant visited rooms in the following order: R1→ R2→ R3→ R4→ R5→ R6→ R7. The estimated sequence of visited locations is: R1→ R2→ R3→ R8→ R4→ R5→ R6→ R7. . . . .* 92
- 4.4 *The location identification performance on transcription of the development dataset as a function of number of LDA topics. The performance on each number of topics presents the average of five experiment with different initial LDA topics. . . . .* 95
- 4.5 *The ASR transcription WERs on different SNRs are shown with a dashed line. The red line shows the WD errors of the LDA-based method for transition detection on the automatic transcription of test data. The black line shows the system performance using the TF-IDF vector representation. . . . .* 97
- 4.6 *The transition detection performance on different transcription WERs. The red line shows the WD errors of the LDA-based method for transition detection on the automatic transcription of test data. The black line shows the system performance using the TF-IDF vector representation. . . . .* 97

4.7	<i>The ASR transcription WERs on different SNRs are shown with a dashed line. The red line presents the LDA-based location identification performance (F1). The black line shows the system performance using the TF-IDF vector representation. . . . .</i>	99
4.8	<i>The location identification performance on different transcription WERs. The red line presents the LDA-based location identification performance (F1). The black line shows the system performance using the TF-IDF vector representation. . . . .</i>	99
4.9	<i>New node insertion by identifying transitions utterances (TU) in the automatic transcription of a spoken report. Each node comprises the utterances of its corresponding (U) segment. . . .</i>	103
4.10	<i>(a) Visualisation of an estimated correspondence matrix (C) for a Map<sub>4</sub> conversation example. Correspondence scores are presented with gray levels ranging from white for zero, as an indication of no match between a pair of nodes, to black for one as a full correspondence. (b) The ground-truth correspondence matrix (GT) of the conversation. . . . .</i>	105
4.11	<i>Visualisation of the estimated correspondence matrix C presented in Figure 4.10a after converting it into its binary form by applying a threshold of 0.5. . . . .</i>	106
4.12	<i>A graphical visualisation of folding a sequence of estimated nodes from places which appear to correspond with each other and transforming it into a likely topological map. . . . .</i>	107
4.13	<i>ROC curves at different SNRs for each map-setting. ROC curves close to the dashed line in the diagonal explain a random estimation.</i>	109
4.14	<i>The ASR transcription WER on different SNRs is shown with dashed line. The AUC for each map-setting is presented as a function of SNR. Random estimation scores a value close to 0.5.</i>	111
4.15	<i>The overall AUC performance of the systems as a function of WER. The red line illustrates performance of the system with automatic segmentation. The black line represents the pre-segmented system performance. . . . .</i>	111

5.1	<i>A general structure of the proposed two-pass speech decoding architecture. The location-ID module provides location estimations from highly inaccurate output of the ASR system. The second pass decoding stage is initiated whenever a new location is identified. All of the stored word lattices related to the recently identified location are then rescored based on the dynamically generated language model. The best path is computed for each word lattice as the final decoding hypothesis. . . . .</i>	115
5.2	<i>The process of building the context-dependent language model in the rescoring pass. The background language model is a large model trained on out-of-domain data (Switchboard corpus transcriptions) and the location-specific language models were small but specific models trained on location-specific collections of utterances in the training data. A context-dependent model was built by dynamically interpolating these static models based on the estimated <math>\lambda_{1-13}</math> coefficients. . . . .</i>	119
5.3	<i>The detailed information about the performance of the baseline system, including its word insertion, deletion, and substitution ratio on different SNRs. . . . .</i>	120
5.4	<i>WER of the baseline ASR system compared with the ELI-condition system as a function of SNR increase. Performance of the location-ID module as a function of SNR is also presented. . . .</i>	121
5.5	<i>(a) The absolute WER reductions for the system in both oracle and ELI-conditions. (b) The WER reductions relative to the baseline ASR for the system in both oracle and ELI conditions. . .</i>	123
5.6	<i>Estimated LMSF on the development-sets for each SNR. . . . .</i>	124



6.1	<i>An envisaged speech decoding architecture. The speech-based localization and mapping module can provide its estimations from a combination of speech and other information sources. These estimations can be used to contribute in updating the locational information gathered during the time of a search and rescue mission. The second decoding pass can use the collected information to dynamically generated a location-specific language model to be used in rescoring the generated word lattices. The best path of each word lattice is the final decoding hypothesis. . .</i>	135
E.1	<i>A typical example of hand drawn topological map of the Map<sub>1</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004). . . . .</i>	188
E.2	<i>A typical example of hand drawn topological map of the Map<sub>2</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004). . . . .</i>	189
E.3	<i>A typical example of hand drawn topological map of the Map<sub>3</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004). . . . .</i>	190
E.4	<i>A typical example of hand drawn topological map of the Map<sub>4</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004). . . . .</i>	191
F.1	<i>The baseline ASR WER landscape at different SNRs and all LMSFs</i>	194
F.2	<i>The baseline ASR performance landscapes at different SNRs and all LMSFs. The ASR performance is normalized between zero and one at each SNR. . . . .</i>	195



# List of Tables

---

2.1	<i>Parameters to characterize ASR tasks, with examples of easy and difficult tasks (Holmes and Holmes, 2001, chapter 15, p. 235).</i>	19
2.2	<i>Performance of various segmentation algorithms on broadcast news data. . . . .</i>	46
2.3	<i>Performance of various segmentation algorithms on dialogue data.</i>	47
3.1	<i>List of the rooms in each map and their types. . . . .</i>	74
3.2	<i>Recording set-up information and the specific recording instruments used. . . . .</i>	76



# Abbreviations

---

**ASR** Automatic Speech Recognition

**AUC** Area Under an ROC Curve

**BEEP** British English Example Pronunciation dictionary

**ELI** Estimated Location Information

**FPR** False Positive Ratio

**FR** First Responder

**GMM** Gaussian Mixture Models

**HMM** Hidden Markov Model

**IDF** Inverse Document Frequency

**LDA** Latent Dirichlet Allocation

**LMSF** Language Model Scaling Factor

**LM** Language Model

**LSA** Latent Semantic Analysis

**LVCSR** Large Vocabulary Continuous Speech Recognition

**MFCC** Mel Frequency Cepstral Coefficient

**NIST** National Institute of Standards and Technology

**PLSI** Probabilistic Latent Semantic Indexing

**PPL** Perplexity

**ROC** Receiver/Operating Characteristic

**SCBA** Self-Contained Breathing Apparatus

**SLU** Spoken Language Understanding

**SNR** Signal-to-Noise Ratio

**SSAR** Sheffield Search and Rescue

**SVM** Support Vector Machine

**TF-IDF** Term Frequency-Inverse Document Frequency

**TL** Task Leader

**TPR** True Positive Ratio

**TREC** Text REtrieval Conference

**WD** WindowDiff

**WER** Word Error Rate

**location-ID** location identification

## Introduction

---

The research presented in this thesis is undertaken as part of a network project ‘*Search and Rescue 2020*’ funded by The University of Sheffield whose aim is to develop novel assistive technologies to enhance and complement the capabilities of humans in search and rescue missions conducted in the year 2020. This network project consists of three interdisciplinary and interlinked projects that brings together researchers from different departments at the University of Sheffield such as Computer Science (COM), Psychology (PSY), Automatic Control and Systems Engineering (ACSE), and Architecture (ARCH). The three network projects, their interrelationship and which of the challenges they address are highlighted in Figure 1.1.

Each of the projects will address a key technological challenge in the area of search and rescue. However, an overarching theme of the project is the development of technologies that aid the overall command and control in search and rescue by providing more accurate and timely sensing, situational awareness and support to the rescue workers. This thesis focuses on ‘*the role of voice communication in command and control*’ as a part of this network project by

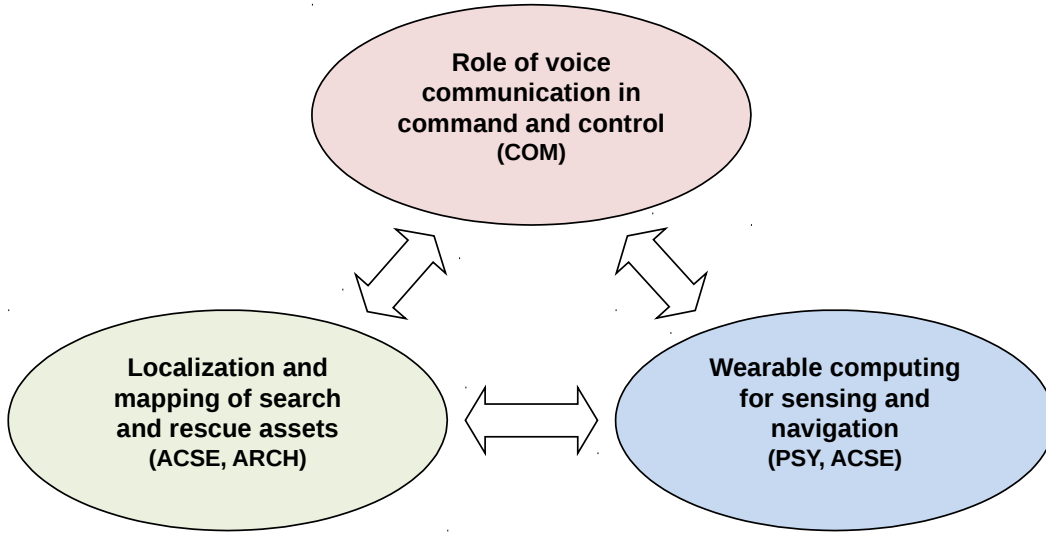


Figure 1.1 Three parts of the ‘*Search and Rescue 2020*’ network project.

investigating the feasibility of developing a system for estimating first responders' location and the incident scene layout by discovering the context and content of voice communication channels.

The rest of this chapter is organised as follows: the motivation that supports the investigations, designs and implementations of this thesis is presented in Section 1.1. The problems that require to be addressed through this research is described in Section 1.2. The principal aim and objectives to be accomplished in this work are presented in Section 1.3. Finally, Section 1.4 presents the organisation of the chapters that build the rest of this thesis.

## 1.1 Motivation

Speech is the primary means for communication between human beings. People use speech to exchange valuable information for perform-



ing their daily routines or accomplishing highly challenging collaborative tasks. Despite the abundance and richness of this source of information, when people talk to each other there is no record and the words are effectively lost. Recent years have witnessed significant improvements in the technology of automatic speech processing. This has led to new interests in both the academic and commercial worlds into the processing of natural spoken conversations and the automatic extraction of their information content. By far the most common place to find such interest is in tracking meetings, analysing customer service calls and extracting valuable information (such as topics discussed, decisions made or customer satisfaction) for management purposes. However recently, attention has been drawn to the role of processing speech communication channels in more critical and challenging application domains such as emergency services (e.g. fire, ambulance, etc.) (Kalashnikov et al., 2009). Techniques are required to help extract and digest situational information for improved decision-making.

The motivation for this thesis comes from a need for automated solutions to extract valuable information from speech communications within the *Search and Rescue* domain. This thesis is concerned with discovering the context and content of communications to provide some form of understanding about an incident scene (discussed here briefly).

In any search and rescue operation, speech is one of the most important sources of situational information. This is the case because, as well as being active participants in the physical aspects of crisis response, human operators also play a central role in the coordination, collection and interpretation of mission-critical information, and the

communication of such information between the relevant parties is mainly achieved using speech.

Speech is widely used for transferring critical information about the incident scene layout and the location of rescue assets. This is because locational information is known to be one of the main enhancing factors for situational awareness formation, and this has direct implications on the efficacy of the response (Shimanski, 2008). For example, a commander in a control centre may issue voice commands (over the radio) to the field operatives instructing them where to move or warning them away from dangerous areas. Similarly, a searcher may report back critical information about their observations, actions, and events at their particular location, or ask for information on where needs to be explored next.

Indeed, these communications and spoken reports can be viewed as verbal annotations of the incident scene. Access to this fast-updating source of information could be vital for the overall performance of the response. In addition, transcribing and saving speech communications are necessary for post-mission analyses including retrieval and browsing of spoken audio documents. Despite their importance, currently human operators are identifying, transcribing and integrating such information into the mission management system *manually*. Whilst individual operators can be highly reliable, the density of voice traffic in the fast moving and dangerous situations may mean that critical information is missed. As a consequence, there is a clear need for the introduction of some form of automation that can either reduce the workload on human transcribers or replace them entirely. Hence, a bright future can be envisaged for crisis response systems which employing speech recognition and understanding systems. Systems can be introduced to help the current support systems for extract-

ing critical information from all conversations. Potentially this can contribute to situational awareness formation for both rescue workers and managers during crisis response.

## 1.2 Problem definition

The importance of automatic extraction of locational information from voice communication channels has been envisaged in the observational-speech-system (Kalashnikov et al., 2009). However, technical difficulties such as, highly imperfect automatic transcription and understanding of conversational and noisy speech, present major challenges for implementing such system.

Although the most advanced *Automatic Speech Recognition* (ASR) systems have now reached the performance of humans on specific datasets and tasks (Xiong et al., 2016), they still have a lot of difficulties in many natural scenarios. In addition to high acoustic variations (mainly caused by environment noise condition, speaker's accent, spontaneous speaking style, etc.), the statistical properties of the language often varies during a conversation due to context change: all of which makes the speech recognition task extremely challenging. Furthermore, conversational speech is not as well-formed as spoken queries directed at a machine, lectures or structured forms. Ungrammaticality and disfluencies, such as false starts, repetitions, and hesitations, are pervasive in conversational speech (Shriberg, 1996). As a consequence, for information extraction from automatic recognition of speech communications, a system needs to be designed which is able to handle these challenges gracefully. Fine-grained identification of fundamental units of meaning (such as sentences, named entities, and dialogue acts) is sensitive to transcription errors and speech dis-

fluencies (Palmer, 1999; Przybocki et al., 1999; Miller et al., 2000). In contrast, topic detection techniques have been reported to be robust to these challenges (Fiscus and Doddington, 2002; Barnett et al., 1997). This leads to the first research question addressed by this thesis: *can topic detection techniques be used to derive high-level information (such as location information) from speech communication channels in a search and rescue environment?*

Post-mission analyses and the retrieval and browsing of spoken audio documents could also suffer from the imperfect transcription of speech communication. In complex speech recognition tasks (such as conversational speech, broadcast news and internet voice-search), high-level contextual information is often used as prior knowledge for guiding the search to determine the most likely sequence of spoken words. This information is commonly gathered from a variety of external sources, for instance, from a mobile phone's geolocation signal (Chelba et al., 2015). The second research question addressed by this thesis is: *can high-level situational information derived from speech communication channels be used top-down to improve speech recognition performance?*

### 1.3 Aims and objectives

The primary aim of this thesis is to investigate the feasibility of developing an automated solution for estimating first responders' location and the incident scene layout by discovering the context and content of voice communication channels. In light of this aim and the above research questions, the following milestones were set as intermediate steps towards this goal:

- Identify the major challenges and limitations in automatic processing of speech communication channels in a search and rescue domain.
- Survey the related background issues and the state-of-the-art in processing natural speech conversations.
- Provide an appropriate speech dataset comprising task-related annotated conversations by targeting the goals and needs of the information extraction task in the context of crisis response.
- Development and evaluation of topic-based locational information extraction in the context of a simulated search and rescue communications.
- Investigate the utility of exploiting the extracted high-level locational information for improving speech recognition performance.

## 1.4 Organization of the thesis

This thesis is structured as follows: Chapter 2 gives a brief introduction to the particular application domain of fire search and rescue voice communication system. This includes a description of the role of speech technology in accessing information content for situational awareness formation. Major challenges and limitations in automatic processing of voice channels are reviewed. To cover the related background issues and state-of-the-art methods in processing conversational speech, Section 2.2 and 2.3 provide a brief overview of speech recognition and understanding systems that are related to the presented research work. This includes a description of the challenges using speech input for understanding tasks and a review of the major topic segmentation and identification approaches in the literature.

The last section of the Chapter 2 provides a short background to automatic topological mapping methods that are utilized in Chapter 4.

Chapter 3 presents a new goal-oriented conversational speech corpus. It starts with a discussion about the necessity of designing and collecting a new speech corpus. Subsequently, the design of a conversation task is described. Finally, the process of dataset collection is described by explaining the recording set-up and annotation scheme.

Chapter 4 introduces an approach for estimating the location of first responders by framing this problem as a topic identification task on their spoken reports. A similar approach is then applied for performing the main steps in a map building technique to interpret such descriptions as a topological representation of the incident scene. After describing the location identification and mapping systems, a set of experiments were carried out on speech data with different environment noise levels and each system performance is reported subsequently.

Chapter 5 investigates the utility of exploiting the high-level location information content of a conversation for improving speech recognition performance. A new two-pass speech decoding architecture is presented. In this architecture, the location estimation from a first decoding pass is employed to dynamically adapt a general language model which is subsequently used for rescoring the initial recognition hypotheses. The recognition performance of the presented system is compared with a baseline speech recognizer by performing a set of experiments on speech data with different noise levels.

Chapter 6 summarizes the thesis, answers to the research questions, describes the limitations of this research and discusses a number of potential directions for future works.



## Summary

*Access to the information content of speech conversations is important for situation awareness formation and the overall performance of any search and rescue operation. Automatic systems can help the current support systems for extracting critical information from all conversations. The scope of this study is focused on extracting valuable situational information from speech communication channels. The primary aim of this thesis is to investigate the feasibility of developing an automated solution for estimating first responders' location and the incident scene layout by discovering the context and content of voice communication channels. This chapter provided an overview of the motivations, problems, research questions, aims and objectives in this research.*





## Background

---

The first part of this chapter (Section 2.1) contextualizes the work presented in this thesis by providing a brief overview of fire search and rescue voice communication system. The role of speech technology in accessing valuable information for situational awareness formation is then presented in Section 2.1.1. Section 2.1.2 describes the major challenges and limitations in automatic processing of voice channels. A brief overview of the speech recognition and understanding systems that are related to the presented research work is provided in Section 2.2 and 2.3. This includes a description of the challenges using speech input for understanding tasks and a review of major topic segmentation and identification approaches in the literature. Finally, Section 2.4 provides a short introduction to automatic topological mapping algorithms that are utilized in Chapter 4.2 for a speech-based topological map estimation in the search and rescue context.

## 2.1 Speech communications in search and rescue

The search and rescue response is a cycle with five different phases: preparedness, mobilization, operations, demobilization and post-mission phase (UN-OCHA, 2012). An effective information exchange during and between all phases results in a coordinated, efficient and, safe response. Speech communication channels play a pivotal role in immediately transferring back the important information following a standardized command hierarchy from first responders to the management team and the task force leader (Schaitberger et al., 2016).

A complete communication map of the crisis response scenario can be so complicated and dense that it is hard to illustrate graphically. This is because it should represent the communications between all components of a response team, including management, search, rescue, medical and logistics. Without paying attention to fine details, Figure 2.1 gives an overview of a typical fire response communication scenario, summarizing several guidelines and reports such as: Schaitberger et al. (2016); UN-OCHA (2012); Wong and Robinson (2004) and NYS-USAR, (2007). This figure also presents the voice and language parameters with a focus on characterizing the difficulty level of the automatic speech recognition task across a fire response process from receiving calls for reporting an incident to the search and rescue teams on the ground (see Section 2.1.2). In most cases, an operation starts by receiving calls for reporting an incident. Based on the caller and the incident location, calls are redirected to specific centres. The assigned centres to the incident work for the local team manager who is located at the base of operation. The base of operation is the focal point of communications and serves as the communications hub on the incident scene. The team manager is responsible for the

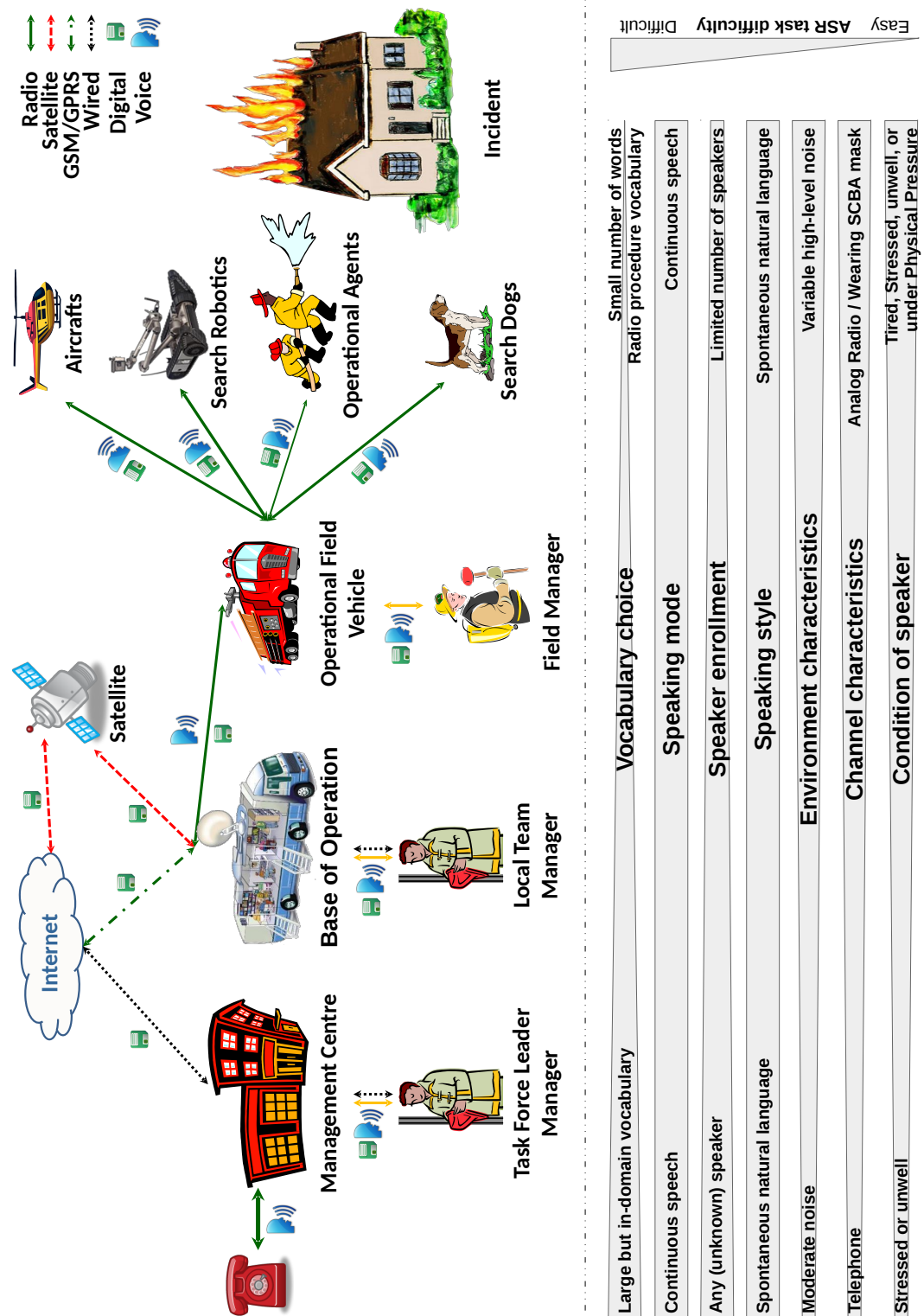


Figure 2.1 top: An overview of a typical fire response communication scenario. bottom: the voice and language parameters at each stage of the fire response process (visualized on top) are presented with a focus on characterizing the difficulty level of the ASR task. The triangle shaped bar indicates the ASR task difficulty at each stage. SCBA stands for Self-Contained Breathing Apparatus.

communication and coordination between various parties involved, as well as other search and rescue teams on the ground. Simultaneously, the team manager documents specific incident scene events, handles requests for additional resources and records tactical radio traffic.

According to the UN-OCHA (2012)<sup>1</sup> statement, all components of every search and rescue team (i.e. management, search, rescue, medical and logistics) should be able to communicate within the team and also with other actors within the theatre of operations via satellite phone, VHF<sup>2</sup>/UHF<sup>3</sup> radio, internet access and/or cellular phones. Even in large-scale sudden-onset disasters, when there is no direct communication between all levels, the commands and reports are normally passed as paper messages to the radio operator who reads them out. In this situation, the management team use predefined tactical symbols on the paper map of the operation site to generate an overview of the rescue operation scene based on the information received. The management team then compiles the final information and reports to the on-site operations coordination centre along with information on the casualties, missing persons and other important information regarding each work-site. In some major natural, technological and environmental disasters, the compiled information and reports from each response team are made available on a ‘global disaster alert and coordination system’ which provides the international disaster response community with near real-time alerts about natural disasters around the world and tools to facilitate response coordination.

In many search and rescue departments, such as the fire service, a priority is given to face-to-face voice communications instead of over

---

<sup>1</sup>United Nations Office for the Coordination of Humanitarian Affairs

<sup>2</sup>Very high frequency range of radio waves (from 30MHz to 300MHz)

<sup>3</sup>Ultra high frequency range of radio waves (from 300MHz to 3GHz)

the radio whenever the information is not needed by the task force leader (Schaitberger et al., 2016). This is due to the limited number of possible radio channels and also difficulty in exchanging complicated and advanced information. However, the sensitive situational awareness, command, and control information still need to be transmitted over radio channels between the members of a crew and their managers.

### **2.1.1 Speech technology for search and rescue**

Speech technology applications in crisis response range from speech enhancement, stress analysis and survivors detection using acoustic listening devices (Wong and Robinson, 2004), to employing automatic speech recognition and synthesis systems which enables firefighters to have a hands and eyes free multimodal communication with a decision support system (Löffler et al., 2006). A comprehensive study on the use of speech and language technology in the context of military environments is presented by Pigeon et al. (2005). Probably the most valuable yet challenging application of speech technology in search and rescue missions is the extraction of mission-critical information from voice channels. Extracting and structuring this information can contribute to situational awareness formation which has direct implications on the efficacy of the response.

#### **2.1.1.1 Situation awareness**

Endsley (1995) introduced and defined situation awareness as the perception of the elements in the environment with respect to time and/or space, the comprehension of their meaning and the projection of their status after some variable has changed, such as time.

In another work, Shimanski (2008) describes situation awareness as “*the degree of accuracy by which one's perception of his/her current environment mirrors reality*”.

One of the primary factors in accidents attributed to human error is related to the lack of or inadequate situation awareness. This is particularly important in complex and dynamic tasks with high information flow, in which poor decisions may lead to serious consequences (e.g. firefighting, air traffic control, military command and control). Walker (1991) explained the lack of coordination in the early years of search and rescue response. He stated that there was “*not only lack of coordination between international teams on the ground, but often the host authorities have no idea of what the specialist teams are capable of, what equipment they have brought with them and often indeed, which teams have actually arrived*”. Thanks to modern technologies, today's firefighting is more coordinated and efficient compared to those days.

The need to access more accurate information from the incident scene automatically and to provide real-time situational awareness for firefighters is the motivation to introduce new technologies to the crisis response system (Löffler et al., 2006). The most important factor for enhancing situation awareness is to increase the reliable and up-to-date resources which are being used in decision-making. Within the search and rescue context, *accessing the incident scene layout* is known to be one of the main enhancing factors in the process of situation awareness formation (Shimanski, 2008).

### 2.1.1.2 Speech-based situation awareness

Mehrotra et al. (2004) and Ashish et al. (2008) considered voice conversations as a source of critical information along with other sources like video data transmitted from cameras, sensor data streams, and textual materials in databases. Employing advanced techniques in speech and language processing can help the current support systems in accessing these conversations. Automatic speech recognition has many potential applications including command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues. Spoken language understanding systems, in particular, offer the potential to enable the current support systems to extract critical information from all conversations automatically, and create situational awareness for both rescue workers and managers during crisis response. Kalashnikov et al. (2009) envisaged an *observational speech system* which can provide these types of situation awareness by observing the human/human communications and understanding the context and content of such communications. However, the limited available studies about processing voice communications in this field barely scratch the surface of its key role in the search and rescue environment.

In some attempts, access to speech data is mainly limited to the retrieval of speech communications from search and rescue mission archives. For instance, Schneider et al. (2007) have designed a system for indexing and retrieving speech data from a search and rescue multimedia archive by recognizing a set of predefined and domain-related keywords. To enhance the robustness and performance of the system, the list of keywords was kept as small as possible to just a limited number of 120 keywords. Stein et al. (2012) have presented an

infrastructure solution for chat transcription for firefighter broadcast communication which relies on the recognition of similar patterns that are frequently used in public safety communications. In the context of urban patrol and reconnaissance, Massie and Wijesekera (2008) have envisaged an interactive voice response service which empowers dismounted soldiers with the ability to access and retrieve tactical information assets from back-end systems using a customized interface based on a small vocabulary continuous ASR system designed for a limited set of users.

### **2.1.2 Challenges in the automatic processing of voice channels**

Processing search and rescue voice communication channels is a significantly complex task for current speech recognition and understanding systems. Section 2.2 reviews the relevant speech recognition technology, but here, Table 2.1 characterizes the main parameters which influence the difficulty of a recognition task. In accordance with the information provided in this table, the previously presented Figure 2.1 shows the voice and language parameters and the difficulties for a recognition task in different areas of the fire response process. For instance, an automatic system for analysing the received calls at the front-end of an emergency service must be capable of handling telephone quality calls from an unlimited number of ordinary citizens. The calls can be reports about different (but limited) situations in a spontaneous speaking style. The particular task of processing first responders' radio communications on the incident scene has its own characteristics. In the following, the main factors that influence auto-



Table 2.1 *Parameters to characterize ASR tasks, with examples of easy and difficult tasks (Holmes and Holmes, 2001, chapter 15, p. 235).*

Task parameter	Easy task	Difficult task
Vocabulary choice	small number of distinct words	unlimited vocabulary or acoustically similar words
Speaking mode	isolated words	continuous speech
Speaker enrolment	known speaker	any (unknown) speaker
Speaking style	read speech, or speech with a strict syntax	spontaneous natural language
Environment characteristics	consistently quiet	variable high-level noise
Channel characteristics	studio quality, close-talking microphone	telephone, with variation in handsets and networks
Condition of speaker	healthy, relaxed and not stressed, but alert	unwell, tired or stressed

matic recognition performance of these communications are described in more detail.

### Environment characteristics:

The environmental noise in a crisis response scenario can fall into the high-noise category (aircraft, factory floor). A variety of high acoustic noises makes this one of the most challenging environments, not only for the automatic speech recognition systems, but also for humans (Schaitberger et al., 2016). There is a variety of noises in this environment. Water bridge engine noise, fire noise and the low-air alarm inside the *Self-Contained Breathing Apparatus* (SCBA) are a few of the most common. Different approaches have been suggested to overcome the noise challenge such as using electromyography-based speech recognition (Betts et al., 2006) or a portable microphone array (Stupakov et al., 2012). Another difficulty, which is also associated with environmental characteristics, is that users often change the way

they speak when the environment changes (the Lombard effect (Lombard, 1911)), for example shouting in extreme fire noise or near a fire engine.

Speech recognition and understanding systems are typically composed of sequential and independent components (see Section 2.3). A challenge for an information extraction (or later a situational awareness system) for crisis response is robustness to high errors in the speech recognition outputs. Kalashnikov et al. (2009) suggested two different yet complementary techniques to overcome the data quality challenge: robustness techniques that exploit a variety of contextual and domain knowledge/semantics to mask errors and improve data quality, and design of data analysis techniques that can tolerate errors in data. They suggested using more semantically enriched representation of the situation like emotion to further process the multiple hypotheses outputs of the speech recognition system.

### **Channel characteristics:**

Radio communication system is a key component of firefighting and fire-ground safety. The form and function of the firefighting radios have not improved much over the past decade (Schaitberger et al., 2016). Although analog radio messaging suffers from bad audio quality, it is still the main channel for information exchange within most response systems. Kushner et al. (2006) reported that human word recognition is lower when using digital radios in comparison to analog radios. He reported that this difference increases when a disturbing factor, such as using a breathing mask, changes the natural utterance and distorts the acoustic signal. Another factor that convinces firefighters to use analog transducer systems is that the analog systems

can carry valuable extra information. For example, when an analog radio user goes into some place with low radio signal coverage, in contrast to digital radios, the acoustic signal gets slowly noisier which gives the user hints before a complete loss of communication Schaitberger et al. (2016). Other disturbing factors like shouting through a SCBA breathing masks into a shoulder-mounted or hand-carried radio, result in a low-quality speech signal. The effect of SCBA masks on the human voice and its intelligibility was investigated by Kushner et al. (2006).

**Physiological/psychological condition:**

Search and rescue operations are often conducted under physiological and psychological conditions induced by high workload, high emotional tension, and other conditions commonly encountered in an incident scene. These conditions are known to affect human speech. For instance, the challenging environment of the incident scene often force the firefighters to communicate while they are crawling on the floor or operating in a face down position. Smoke or toxic gases can change an individual's voice characteristics. Other relevant factors, including fatigue, emotional and physical stress, can also change the speaker's voice and present problems to speech processing equipment such as voice coders, automatic speech and speaker recognition systems. Similar challenging conditions are also identified in the context of military communication (Pigeon et al., 2005).

**Speaking style:**

The natural spontaneous speaking style of conversations may not be grammatical. These conversations often include a large number of

disfluencies (such as hesitations, errors and corrections, mispronunciations, etc.) which is generally harder to recognize than read speech.

On the other hand, recognition performance can be higher compared to the task of transcribing everyday speech. First responders are trained to follow a strict standardized communications procedures with small distinct terminologies such as voice/radio procedures and vocabulary (NFPA, 2014). An automatic speech recognition system can also be speaker dependent and adapt to a speaker during training for achieving higher recognition performance.

## 2.2 Automatic speech recognition

Spoken language processing systems are typically composed of sequential and independent components. Automatic recognition of speech is often the starting processing stage for further components such as *Spoken Language Understanding* (SLU) systems (Tür and De Mori, 2011). The purpose of an ASR system is to convert an incoming speech signal into the most likely word sequence. Figure 2.2 shows the general architecture of an ASR system and its main components. A brief introduction to each component is provided as follows.

### Front-end processing:

Input speech signals are preprocessed in the first stage of speech recognition. This initial stage, which is usually known as front-end processing or feature extraction, provides a stream of fixed size *acoustic feature vectors*, or *observations*  $O = o_1, o_2, \dots, o_t$ . The front-end processing aim is to extract compact observations for the recognition task. Two widely used speech feature representations in state-of-the-art speech recognition systems are *Mel-Frequency Cepstral Coeffi-*

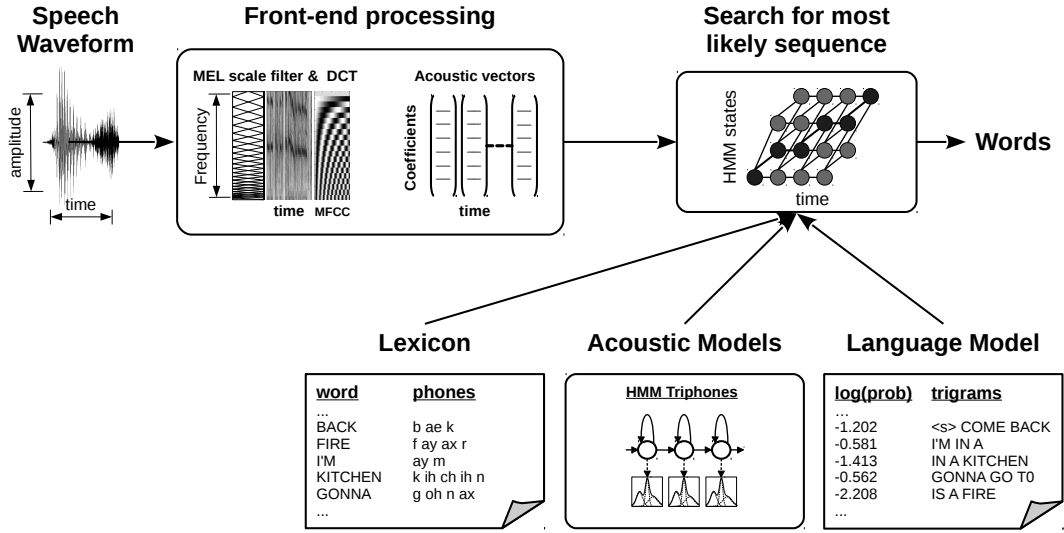


Figure 2.2 A general architecture of a large vocabulary continuous speech recognizer.

coefficients (MFCC) (Davis and Mermelstein, 1980) and *Perceptual Linear Prediction* (PLP) coefficients (Hermansky, 1990). In some application domains, such as telephone speech recognition or broadcast news transcription, the front-end also isolates relevant speech segments from the whole audio stream in a process called *segmentation*.

## Decoding:

In the second stage, the extracted sequence of acoustic feature vectors is fed into a decoder (also known as search or inference component) to recognize the sequence of words  $W = w_1, \dots, w_n$  (each one drawn from a vocabulary  $\mathcal{V} = v_1, v_2, \dots, v_V$ ) which is most likely to have generated  $O$ . More formally, the decoder tries to find:

$$\hat{W} = \operatorname{argmax}_{W \in \mathcal{V}} P(W|O) \quad (2.1)$$

Although  $P(W|O)$  can be modelled directly using discriminative models, Bayes' rule is used to transform equation 2.1 into the equivalent

problem of finding:

$$\hat{W} = \operatorname{argmax}_{W \in \mathcal{V}} \frac{P(O|W)P(W)}{P(O)} \quad (2.2)$$

Since  $P(O)$  is fixed acoustic evidence and does not change over the recognition process, the search problem can be decomposed into two parts, the *acoustic modelling problem*, and the *language modelling problem* (Rabiner and Juang, 1993, chapter 8):

$$\hat{W} = \operatorname{argmax}_{W \in \mathcal{V}} \underbrace{P(O|W)}_{\text{Acoustic Model}} \underbrace{P(W)}_{\text{Language Model}} \quad (2.3)$$

Two commonly used decoding algorithms are time-synchronous *Viterbi* decoding, which is usually implemented with pruning and can then be called ‘*beam search*’, and stack or A\* decoding (Jurafsky and Martin, 2009, chapters 9-10). Given a sequence of cepstral feature vectors as input, ASR systems mainly recognize the sequence of words based on three main knowledge sources: i.e., *lexicon*, *acoustic models* and, *language models*.

### Lexicon:

The *lexicon*, also known as the *pronunciation dictionary*, is simply a list of words, with a pronunciation for each word expressed as a phone sequence. While most words have a single pronunciation, some words may have more. The lexicon is used to map phones (the basic unit of sound) to words used in the language model. Publicly available lexicons like the *British English Example Pronunciation (BEEP)* dictionary (Robinson, 1996) with about 250,000 words or the *Carnegie Mellon University American English pronunciation dictionary* (CMU, 1998) with about 64,000 words, can be used for building ASR systems.

**Acoustic model:**

The acoustic model represents the acoustic knowledge of how an observation sequence can be mapped to a sequence of phones. This acoustic knowledge is used to determine the likelihood of  $P(O|W)$ . Since mid 1980's, almost all modern *Large Vocabulary Continuous Speech Recognition* (LVCSR) systems have a *Hidden Markov Model* (HMM) in their acoustic model core to construct the temporal structure of speech (Rabiner and Juang, 1993, chapter 8). In LVCSR, HMMs are normally used to model sub-word units (i.e. monophone or triphone models). Sub-word units are then composed to form word HMMs according to rules specified by the lexicon. Detailed description about the widely used HMM-based ASR systems with *Gaussian Mixture Models* (GMMs) as the state emission can be found in most speech processing textbooks such as Jurafsky and Martin (2009) and Gales and Young (2007). In parallel with the GMM-based systems, various approaches using *Deep Neural Networks* (DNNs) (e.g. Seide et al. (2011); Yu and Seltzer (2011) among many) have become popular for delivering better performance than the GMM/HMM systems on a number of tasks (Hinton et al., 2012).

**2.2.1 Language model**

A crucial and indispensable component of an ASR system is its *Language Model* (LM). The LM guides the search to determine the most likely sequence of words by quantifying the validity of acceptable word sequences in a given language for a given task domain (Rabiner and Juang, 1993, chapter 8). The LM represents the prior knowledge,  $P(W)$ , about the syntactic and semantic information of word sequences. The LM can also improve the recognition performance by

providing contextual information. The challenge in language modelling is to encapsulate as much as possible of the syntactic, semantic, and pragmatic characteristics of a language in a particular task.

Due to the complexity of natural language, it is almost impossible to construct language models using a set of linguistic rules. For this reason, statistical language modelling (such as n-gram modelling and, recently, recurrent neural network language modelling) is the dominant approach over the last few decades. In principle, a statistical model is built from a large amount of training data coming from the same population as a target domain in which the model to be applied.

### 2.2.1.1 N-gram model estimation:

N-grams are the most popular language models employed in the speech recognition task. For a given sequence of words ( $W=w_1, \dots, w_n$ ), the language model estimation using the chain rule and order-2 Markov assumption leads to:

$$P(w_1, \dots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\dots P(w_n|w_1, \dots, w_{n-1}) \quad (2.4)$$

$$\approx P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\dots P(w_n|w_{n-2}, w_{n-1}) \quad (2.5)$$

$$\approx P(w_1)P(w_2|w_1) \prod_{i=3}^n P(w_i|w_{i-2}, w_{i-1}) \quad (2.6)$$

The first two terms in equation 2.6 are called a unigram and a bigram, respectively. The last one is called a trigram since two previous words (i.e.  $w_{i-2}$ , and  $w_{i-1}$ ) are used for conditioning. Higher order n-grams are possible such as 4-grams, 5-grams. The n-gram model estimation is performed using simple maximum likelihood estimates from the



training set data:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{N(w_{i-2}, w_{i-1}, w_i)}{N(w_{i-2}, w_{i-1})} \quad (2.7)$$

$$P(w_i|w_{i-1}) = \frac{N(w_{i-1}, w_i)}{N(w_{i-1})} \quad (2.8)$$

$$P(w_i) = \frac{N(w_i)}{\sum_{w_i \in \mathcal{V}} N(w_i)} \quad (2.9)$$

where  $N(w_{i-2}, w_{i-1}, w_i)$  denotes the number of times word sequence  $\{w_{i-2}, w_{i-1}, w_i\}$  appears in the training data.

N-gram models are easy to build but a problem arises when an n-gram is encountered during testing that was not seen in the training examples. In this condition,  $N(w_{i-2}, w_{i-1}, w_i)$  is zero in the training set, however it is non-zero in the test set. Thus it is necessary to estimate the probability of unseen events in the training set. Therefore, a language model is modified through a process called smoothing so that no word sequence gets zero probability. The basic idea is to reserve (subtract) some small probability mass from the relative frequency estimates of the probabilities of seen examples, and to redistribute this probability to unseen ones. Proposed methods differ according to the reserve value (discounting) and how it is redistributed (back-off). *Kneser–Ney smoothing* is one of the widely used methods in state-of-the-art systems (Kneser and Ney, 1995). More details are seen in research presented by Chen and Goodman (1996), Kim et al. (2001), among others.

### 2.2.1.2 Language model adaptation

Thanks to the exponential growth in the amount and diversity of data available online, the quality of statistical language models has increased. Nevertheless, this tendency appears to be reaching an

upper limit (Rosenfeld, 2000; Bellegarda, 2004) and it is possible that the data increase does not lead to any significant improvement in language models. The statistical language models used in state-of-the-art ASR systems are often trained on more data than a human being ever be able to hear and read in a lifetime (Evermann et al., 2005). For example, Chelba et al. (2015) trained a 5-gram model on about 695 billion words of training data for a voice-search task. However, when tasks become a little more complex (e.g. conversational speech or broadcast news), even the best ASR system's performance declines significantly. This performance decrease is often associated with the fact that the statistical language models are extremely brittle to the highly heterogeneous nature of natural speech, with varying domains, genres, and styles (Rosenfeld, 2000; 1995).

In order to reflect the changes that a language experiences when moving towards different domains, language model adaptation techniques (Bellegarda, 2004) are often used in ASR systems. The adaptation goal is to enrich previously trained models by adding new sources of information to maintain an adequate representation of the task context under changing conditions. A variety of adaptation approaches has been introduced for different statistical language modelling strategies. The basic idea which most of them have in common is the incorporation of some dynamic informative features (either about the discourse context or about the data domain) in the process of training the language model. This has been shown to be effective in many multi-topic tasks such as multi-genre broadcast speech recognition as shown in research by Chen et al. (2015; 2003).

A range of informative features has been employed for language model adaptation. Among these, information derived from analysis of the speech data was utilized in some approaches. For instance,

Chen et al. (2001) used keyword information for model adaptation in the task of broadcast news transcription. Shi (2014) used utterance length and lexical features on lecture transcripts. More robust techniques in the field of information retrieval, such as *Latent Dirichlet Allocation* (LDA) document modelling (Blei et al., 2003) were used by Chien and Chueh (2011), Mikolov and Zweig (2012) and Echeverry-Correa et al. (2015).

Some adaptation approaches are based on the specific context of the task that they are addressing. New sources of information are often used to generate a context-dependent language model. These new sources may come, for instance, from geolocation signals obtained from mobile phones in voice search tasks (Chelba et al., 2015; Halpern et al., 2016), from personalized user information such as demographic features in a social media task (Wen et al., 2013) or from speaker identification systems (Nanjo and Kawahara, 2003).

Other approaches are based on the analysis and extraction of semantic information that the user provides. For example in a spoken dialogue system, dialogue concepts inferred by the dialogue manager, and represented as dialogue goals, were used to adapt the language model (Lucas-Cuesta et al., 2013).

Adding these auxiliary features allows language models to exploit commonalities and specialities among diverse data better. Later at test time, it facilitates the adaptation to any target domain defined by some prior high-level contextual information (often obtained from a variety of knowledge sources) or the characteristics of the test data (e.g by identifying the topic or set of topics). For n-gram language modelling, the adaptation proceeds generally in the following steps:

1. Making subsets of the training corpus based on some informative features (e.g. Iyer and Ostendorf (1999)).
2. Using these subsets to build multiple domain-specific models.
3. At test time, obtaining contextual information such as locational information (Chelba et al., 2015) or identifying the characteristics of the test data such as topic for instance in (Seymore and Rosenfeld, 1997; Seymore et al., 1998).
4. Identifying the relevance of each subset-wide model using the obtained information.
5. Combining the models according to their relevance (via linear interpolation) and making a mixture model of all the domain-specific models.

### 2.2.1.3 Model interpolation

Model interpolation is the most known and widespread strategy for adapting a background model to a more specific domain. This section presents basic idea together with the definitions that have been used in this thesis. Language model interpolation consists of taking a weighted sum of the probabilities given by the component models. Let  $P(w|h)$  be the probability of observing the word  $w$  given the previous sequence of words in its history  $h$ . Given a background model  $P_B(w|h)$  and a domain-specific adaptation model  $P_A(w|h)$ , the final model  $P(w|h)$  can be obtained as:

$$P(w|h) = (1 - \lambda)P_B(w|h) + \lambda P_A(w|h) \quad (2.10)$$

where  $0 \leq \lambda \leq 1$  serves as the interpolation coefficient.

A very common language model adaptation case is when only a small amount of data is available in the target domain and large amounts in other domains. In this case, the in-domain model is combined with the background model via linear interpolation. The interpolation coefficient  $\lambda$  is commonly tuned by minimizing the *Perplexity* (PPL) (Jelinek et al., 1977) on some held-out data similar to the target domain (validation or development dataset).

A generalization of this linear interpolation (equation 2.10) is used to include several predefined domain-specific language models (Bellegarda, 2004). The mixture model probability of  $K$  domain-specific models ( $P_k$ ) is obtained as:

$$P(w|h) = \sum_{k=1}^K \lambda_k P_k(w|h) \quad (2.11)$$

where  $\lambda_k$  denotes the interpolation coefficient of  $k^{\text{th}}$  domain-specific model. Different sources of information can be used to determine optimum  $\lambda$  coefficients for interpolation of  $K$  language models in a given task. These approaches have often been used to dynamically adapt a background model based on some information about changes in the speech domain (Echeverry-Correa et al., 2015; Chelba et al., 2015). Figure 2.3 shows the widespread model interpolation set-up for making dynamic language models. The dynamically adapted language models are generally used in a *second stage decoding* process on the *lattice* output of the initial recognition pass. This process is called *language model lattice rescoring* which is explained in Section 2.2.3.

### 2.2.2 Performance of speech recognition systems

Intensive efforts from the 1980's onwards have improved the development of discrete word, speaker dependent large vocabulary ASR

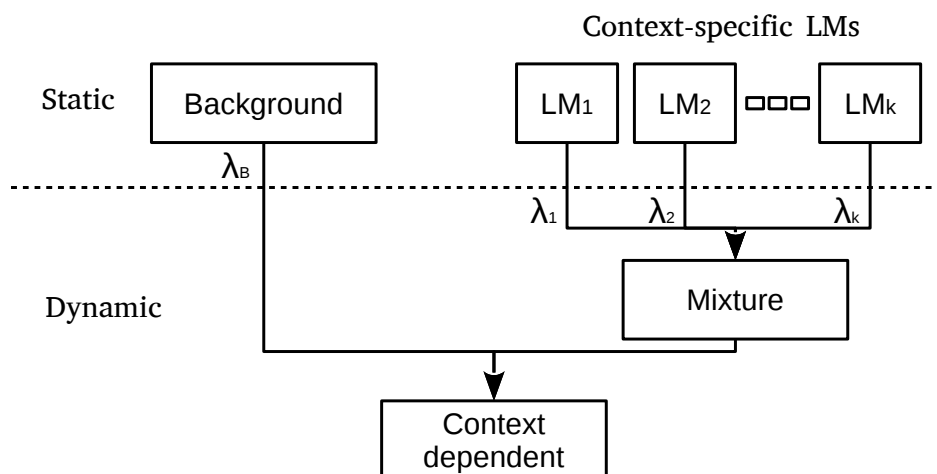


Figure 2.3 *The process of making a context-dependent language model in the rescoring pass. The background language model is typically a large model trained on out-of-domain data and the context-specific language models are typically small but specific models trained on subsets of the training corpus. A context-dependent model can be made dynamically by interpolating these static models based on some information about changes in the speech domain.*

systems and provided a strong backbone to the modern continuous speaker-independent ASR systems. Current fast decoding algorithms allow continuous-speech recognition of large vocabulary sizes in near real-time response. However, none of them are 100% correct. Thanks to the increased computation power provided by high-performance computing systems and recent achievements in employing deep learning in speech recognition, the most advanced ASR system by far has just managed to reach the performance of humans in recognizing conversational speech (Xiong et al., 2016). However, still there is a gap between the recognition accuracy of current ASR systems and the accuracy of humans in recognizing speech in many real life scenarios. This gap becomes more profound when it comes to speech recognition in noise, with channel variability, spontaneous speech and also little contextual and grammatical information.

The standard evaluation metric for speech recognition systems is the *Word Error Rate* (WER). The WER is an intuitive direct measure of how much the hypothesized word string returned by the recognizer differs from a correct or reference transcription. The first step is to compute *minimum edit distance* in words between the hypothesized and correct strings by finding the minimum number of word *substitutions* (S), word *insertions* (I), and word *deletions* (D) necessary to map between them. The WER is then defined as follows<sup>4</sup>:

$$\text{WER} = 100 \times \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Correct Transcript}} \quad (2.12)$$

An example is presented in Figure 2.4 showing a reference, a hypothesis, and each word error types. The WER of this example is:

$$\text{WER} = 100 \times \frac{1 + 1 + 1}{7} = 42.8\%$$

Reference:	IT'S	DIFFICULT	TO	***	DESCRIBE	THIS	ROOM
Hypothesis:	IT'S	DIFFICULT	TO	THIS	RIGHT	***	ROOM
Error:	C	C	C	I	S	D	C

Figure 2.4 An example showing a reference utterance transcript, its ASR hypothesis, and each word error types. C, I, D, and S stand for correct, insertions, deletions and substitutions, respectively.

The standard statistical test for comparing ASR systems performance is the *Matched-pairs* test of their word error (Gillick and Cox, 1989). This is a parametric test that looks at the difference between the number of word errors the two systems produce, averaged across a number of segments. While the implementation of the WER and often *Matched Pairs Sentence-Segment Word Error* (MAPSSWE) test

<sup>4</sup>Since the equation includes insertions, the error rate can be greater than 100%

is available on most speech recognition toolkits such as Kaldi ASR (Povey et al., 2011), the standard implementation of them is provided by the *National Institute of Standards and Technology* (NIST) as a free script called *sclite* (NIST, 2016).

### 2.2.3 Multipass decoding

HMM decoders mainly use the *Viterbi* (Viterbi, 1967) decoder which is a dynamic programming algorithm for finding the most likely sequence of hidden states (called the Viterbi-path) and generate that word string (Gales and Young, 2007). The Viterbi algorithm computes an approximation of the sequence of words which is most probable given the input acoustics. The accuracy of this approximation decreases when each word in its lexicon has multiple pronunciations. A further problem with the Viterbi algorithm is that it is impossible or expensive to incorporate more advanced language models or other high-level knowledge sources for increasing the decoding accuracy (Jurafsky and Martin, 2009, chapter 10).

One solution is to modify the Viterbi algorithm (such as *n*-best algorithm of Schwartz and Chow (1990) among other methods) in a way to return multiple potential utterances instead of a single Viterbi-path by using general and efficient knowledge sources. These multipass decoders generally produce output in the form of *n*-best word string hypotheses, each of which is annotated with an acoustic model probability and a language model probability. A more sophisticated representation is often used called a *word lattice* (Murveit et al., 1993; Aubert and Ney, 1995), which is capable of efficiently representing more information about possible word sequences. Each word hypothesis in a lattice is augmented separately with its acoustic model like-



likelihood and language model probability. Later, during a *second-pass* decoding, another high-level language model or a more sophisticated acoustic model can be used to rescore and re-rank the hypotheses. Figure 2.5 shows a modified form of the standard HMM-based ASR architecture in Figure 2.2 using the multipass decoding strategy.

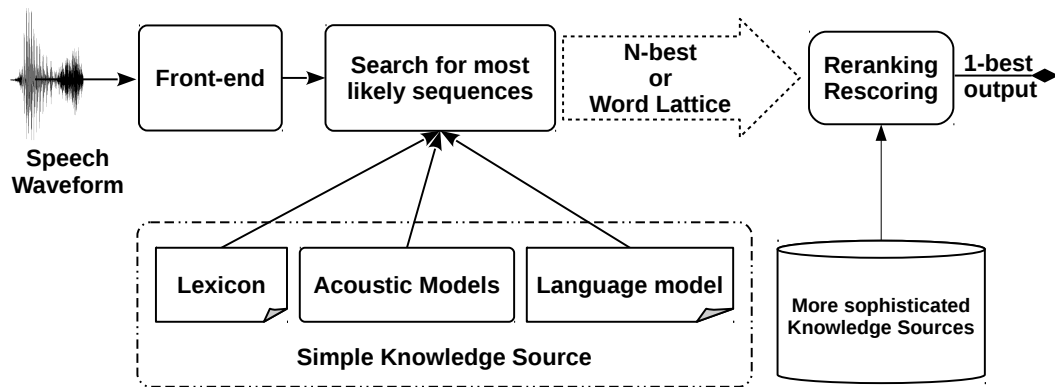


Figure 2.5 The general architecture of a two-stage speech decoding. High-level language model or a more sophisticated acoustic model can be used in a second-stage decoding to rescore and re-rank the initial decoding hypotheses.

### 2.2.3.1 Language model lattice rescoring

A word lattice is a rich structure used to represent a great number of alternative ASR hypotheses in a compact graph form  $G(N, A)$ , where the timing information is embedded into the nodes  $N$  and the arcs  $A$  carry the symbol along with the information about the language model and acoustic model *scores*. This weighted, directed and acyclic graph is typically generated from a single-pass decoding of speech signals using relatively simple knowledge sources. This graph can include multiple word string hypotheses for a given spoken utterance.

*Lattice rescoring* refers to a second decoding pass, over the initial ASR lattice output, with the help of more advanced knowledge sources. A domain-specific language model or a 5-gram model with

a higher order of knowledge can be used to rescore a word lattice. In practice, this is carried out by initially decoupling the scores of acoustic and language models of the transition arcs and then replacing the previous language model probabilities with the *new* model probabilities. In other words, the rescoring process needs to subtract the old model probabilities and then add in the new model probabilities. Several approaches either based on offline or on-the-fly (Sak et al., 2010; Hori et al., 2007) composition have been presented for handling lattice rescoring in extremely large vocabularies. Given additional knowledge sources, include high order n-grams (Hain et al., 2012), location specific language models (Chelba et al., 2015) and contextual articulatory knowledge (Li et al., 2005; Siniscalchi et al., 2006), lattice rescoring methods resulted in higher recognition performances compared with conventional single pass decoding.

## 2.3 Understanding speech conversations

Conversation is the most natural and efficient way for humans to exchange information and coordinate their activities. Two-party conversations, multiparty meetings, and lectures are various types of conversational set-ups which exist in everyday human life. Whilst spoken language understanding mainly refers to the understanding of voice enquiries to personal assistants, interpreting human/human voice communications and integrating the outcomes with relevant information sources is an emerging demand in a variety of application domains. From meeting conversations in a company to speech communications during a crisis response (Kalashnikov et al., 2009), automatic information extraction from these spoken interactions can provide valuable situational knowledge for making better decisions.

The aim in SLU is to extract meaning from natural language utterances. Designing a general purpose framework for tracking, extracting and representing conversation content is difficult (Gokhan Tür and Hakkani-Tür, 2011). Therefore, in practice, SLU approaches tend to depend on the specific intended application area. However, research in the areas of *dialogue act segmentation and tagging*, *named entity recognition*, *topic segmentation*, and *identification* provided a task-independent basis for further discourse analysis and understanding. For example, it has been shown that dialogue acts can be used in practical high-level tasks of extracting key information related to action items and decisions (Morgan et al., 2009) or hot spot detection (Wrede and Shriberg, 2003) in conversations. Named entity recognition is a crucial component of spoken information extraction systems (Makhoul et al., 2000) for the task of performing complex search queries on large audio archives. Topic segmentation and identification is an essential step in understanding and information retrieval tasks. Topic segmentation is often used to divide a long uninterrupted transcript of a business meeting or a news broadcast into shorter and topically coherent segments. By analysing or classifying the contents of each segment, topics from one meeting to another can be related. Similarly, topic detection can be used to track the progress of news stories across different broadcasts, produce a summary with the main headlines of a news story, or the final decision and action items of a meeting.

### 2.3.1 Challenges using speech input

Dealing with speech, more specifically spontaneous speech, the first challenge SLU approaches face is speech disfluencies: hesitations, filled

pauses, false starts, etc. These occur frequently in highly spontaneous speech, such as:

- E'ER IT LOOKS LIKE THE <PAUSE> THE HOB'S CAUGHT FIRE
- AND <PAUSE> IN THE CORNER THERE'S A MAN NO A WOMAN  
ON THE FLOOR

The second challenge of using speech input is to deal with highly imperfect automatic transcription of natural conversational speech which is often contaminated with background noise. The ASR transcriptions contain errors: words can be deleted, replaced or false detection can insert erroneous words. Speech disfluencies and transcription errors represent challenges to the fine-grained identification of the fundamental units of meaning (e.g. sentences, named entities, and dialogue acts) (Gokhan Tr and Hakkani-Tr, 2011). The *Out-Of-Vocabulary* (OOV) word phenomenon in speech recognition is another source of errors which is particularly important in the task of named entity recognition from speech input.

Several studies have shown that named entity recognition performance is strongly correlated with WER (Palmer, 1999; Przybocki et al., 1999; Miller et al., 2000, among others). Miller et al. reported 0.7 points of F-measure<sup>5</sup> lost for each additional 1% of WER. Significant performance drop has also been reported on dialogue act segmentation and identification (Ang et al., 2005).

The impact of recognition errors on the overall performance of topic detection systems have been studied in the NIST *Topic Detection and Tracking* (TDT) (Fiscus and Doddington, 2002) and *Text REtrieval Conference* (TREC) document retrieval (Barnett et al., 1997)

---

<sup>5</sup>The F-measure is one of the evaluation metrics used in named entity recognition performance measurements.

evaluation programs. In contrast with named entity or dialogue act detection, topic detection results presented by Fiscus and Doddington have shown that this impact was very limited. Similar results were also presented by Barnett et al. for a document retrieval task during the TREC program. Hazen (2011) describes the main explanation for this phenomenon as the redundancy effect. Topics are often represented by many occurrences of salient words characterizing them. Even if some of these words are missed or replaced, information retrieval methods can use the remaining informative words and phrases, discard the noise generated by the automatic transcription module. This phenomenon is not true for tasks related to the extraction of fine-grained units of meaning (Frederic Bechet, 2011). The next two sections describe the task of topic segmentation and identification in the domain of speech conversation.

### **2.3.2 Topic segmentation**

Topic segmentation is used to divide a complete recording or transcript into shorter, topically coherent segments. For spoken data, the segmentation is often a necessary first step before topic identification and other deeper processing tasks. Topic segmentation has been used for improving browsing or searching for a particular story in broadcast news (Allan et al., 1998; Doddington, 1998). In another domain, segmentation has been used to aid searching and accessing university lecture recordings, e.g. the Lecture Browser project at the Massachusetts Institute of Technology (Glass et al., 2007), or in the European LECTRA and CHIL projects (Trancoso et al., 2006; Fügen et al., 2006). In the conversation domain (such as business meetings), where the data can be long and involve several topics, indexing by

topic segment has been used to help a user to browse and search for a record effectively (Banerjee et al., 2005; Lisowska, 2003).

### 2.3.2.1 Segmentation techniques

Topic segmentation has been approached in many different ways and most of them share two basic insights, either individually or in combination. The first insight is that a topic change is associated with the *introduction of a new vocabulary* (Youmans, 1991). This is because when people talk about different topics, they discuss different sets of concepts and they use words relevant to those concepts. The second basic insight is that there are *distinctive boundary features* between topics. This is mainly because of the fact that the speaker tends to signal to the audience about switching from one topic to another by using various words/phrases (e.g. ‘*Okay*’, ‘*Now*’, ‘*So*’, ‘*Anyway*’, *etc.*) or prosodic cues (Grosz and Sidner, 1986; Hirschberg and Litman, 1993; Hirschberg and Nakatani, 1998). The advantage of using these boundary features is that they are generally independent of the subject matter and they can be used to estimate the boundaries more accurately in comparison to content-based techniques (Purver, 2011).

Different approaches have been introduced both for content-based and boundary-based segmentations. Hearst introduced the *TextTiling* system (Hearst, 1997; Hearst and Plaunt, 1993) which was one of the early algorithms proposed to use a similarity measure for segmenting broadcast news. It was inspired by classical approaches in the information retrieval domain. In the TextTiling system, the discourse is divided into windows of a fixed width (after some text preprocessing like tokenization). Moving the window across the discourse, each window is represented by a lexical frequency vector. The similarity

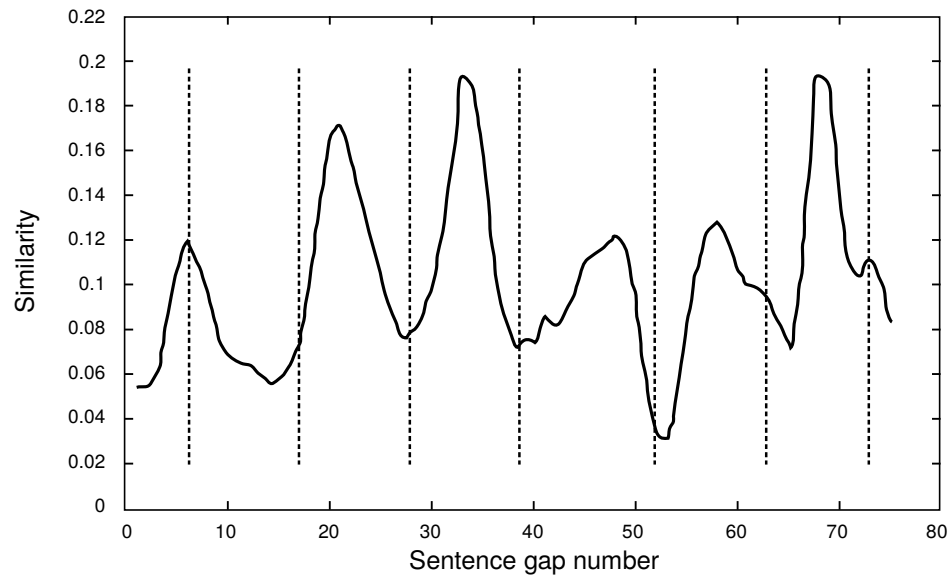


Figure 2.6 *TextTiling* graph. Vertical lines indicate actual topic boundaries. The graph indicates computed similarity of adjacent windows of text. Peaks indicate coherency, and valleys indicate potential breaks between tiles (reproduced with permission from Hearst and Plaunt (1993)).

of each pair of adjacent windows is then calculated using the cosine distance between their lexical frequency vectors. Significant local minima in the lexical cohesion (i.e. the smoothed similarity curve) were considered as an indication for hypothesized topic boundaries (see Figure 2.6). Using the same overall approach, advanced text vectorization techniques have been employed by (Claveau and Lefèvre, 2015) for comparing the similarity between the two windows of text.

*DotPlotting* (Reynar, 1998) and C99 (Choi, 2000) both used clustering on the similarity matrix between candidate segments. To decide if the topic has changed or not, the DotPlotting-based approach relies on word repetition for computing some kind of similarity. The similarity matrix is made by plotting discourse as a two-dimensional matrix with its words along both axes in linear order (see Figure 2.7). A dot (i.e. non-zero entry) is placed wherever words match. Squares can be seen corresponding to topics in areas with more frequent near-

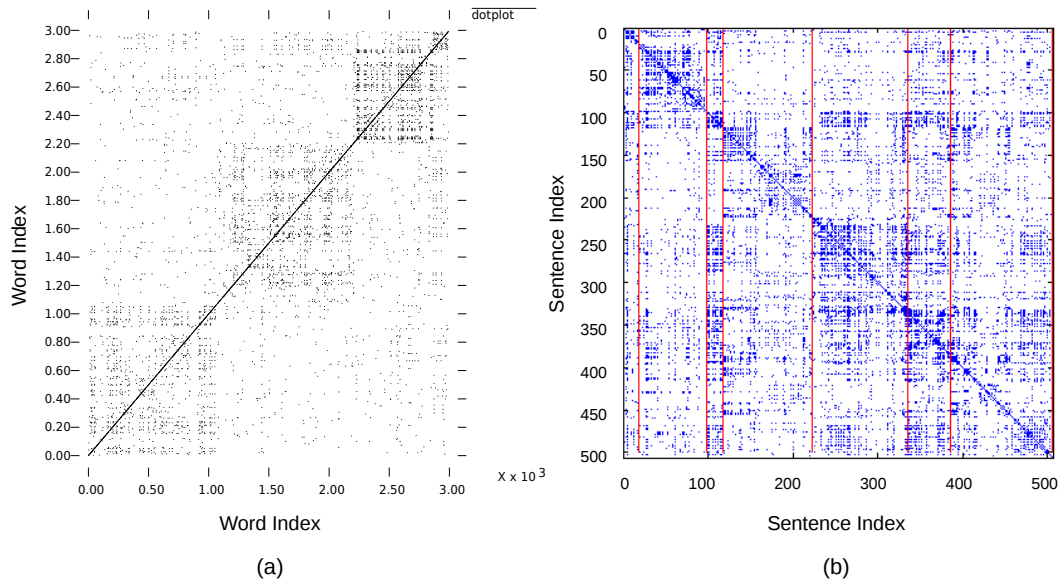


Figure 2.7 a: Dot-plotting of four concatenated Wall Street Journal articles (reproduced with permission from (Reynar, 1994)). b) Utterance similarity plot for a Physics lecture, with vertical lines indicating true segment boundaries (reproduced with permission from (Malioutov and Barzilay, 2006)).

neighbour matching. The topic boundaries are estimated by identifying the boundaries between these squares. The best set of boundaries is estimated in an unsupervised way by maximizing the dot density within the squares, and minimizes the density outside them. Assuming a known number of boundaries, Reynar (1994) used a best-first search algorithm to minimize the density outside the squares.

The same content-based phenomenon has also been exploited from a generative perspective. Yamron et al. (1998); Blei and Moreno (2001) and Purver et al. (2006), among others, used topic language models and variants of the hidden Markov model to identify topic segments. These systems can model the sequence of words as being generated from some underlying sequence of topics which each has its own characteristic word distribution. Making these systems requires a segmented training dataset to estimate the topic language models



and the topic transition probability. However the learning process is unsupervised (based on clustering).

In contrast to content-based approaches, a different strategy has been proposed to look for the characteristics of boundary features. Passonneau and Litman (1997) showed that the cue phrases and the prosodic features that people often use to signal topic change at the beginnings and the ends of topics could all be useful in segmentation. Different domains can have their own specific cue phrases in addition to the general cues such as ‘*So*’, ‘*Anyway*’, etc. For instance Maybury (1998) describes, in broadcast news, phrases such as ‘*Joining us*’, ‘*Tonight*’ and ‘*Welcome back*’ are strongly indicative of topic change. Such features are often automatically learned from labelled training data and used in discriminative approaches such as a *Support Vector Machine* (SVM) classifier to identify topic segment boundaries (Georgescul et al., 2006b; 2007). SVMs are based on the idea of finding a hyperplane that best divides a dataset into two classes. A Linear SVM finds a hyperplane that best separates the data points in the training set by classes label. The hyperplane is called the decision boundary, and cuts the space into two halves one for each class. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class. In general, the larger the margin the lower the generalization error of the classifier. The data often need to be mapped into higher and higher dimensions in a process called Kerneling until a hyperplane can be formed to segregate it. A new point can simply be classified by identifying which side of that hyperplane the point is. Note that this description only applies to binary classification problems and if a dataset has more than two classes, there are other SVM approaches (such as one-versus-all or one-versus-one). Having access to labeled

training data, supervised boundary detection has been reported with higher accuracy (Georgescu et al., 2007; Galley et al., 2003) compared with unsupervised segmentation approaches (see Table 2.3).

### 2.3.2.2 Evaluation metric

The recall and precision metric has often been applied to the topic identification problem (Hazen, 2011). However, because of the nature of segmentation, standard evaluation metrics in classification tasks are not always suitable. In contrast to the identification task, here there is no correct/incorrect answer to be able to count up the scores. Therefore, different scores have been proposed for the segmentation task.

Since recall and precision can be inconsistent without any preprocessing, Beeferman et al. (1999) proposed the  $P_k$ -score, which has been widely used.  $P_k$  expresses a probability of segmentation error, with higher  $P_k$  meaning a less accurate segmentation and a higher probability of error. However, Pevzner and Hearst (2002) have shown that the  $P_k$ -score suffers from some failures in some conditions such as: 1) penalizing missing boundaries more than false alarms; 2) heavily penalizing near-miss errors in comparison to false alarms and missing boundaries; 3) not detecting new segments with size smaller than  $k$ ; and 4) cannot be interpreted as an error percentage (Pevzner and Hearst, 2002). Based on that, Pevzner and Hearst proposed *WindowDiff* (WD) which is usually preferred for evaluating segmentation systems. The WD can be seen as an error rate, with lower WD scores indicating better segmentation accuracy. The WD is calculated by taking a window of fixed width  $k$  and sliding it across the dataset.

At each step, the difference between the number of hypothesized and reference boundaries within the window is counted and the WD score is the average difference values for all windows. It is defined as:

$$WD(ref, hyp) = \frac{1}{N - k} \sum_{i=1}^{N-k} |b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| \quad (2.13)$$

where  $b(x_i, x_j)$  represents the number of boundaries between  $i^{th}$  and  $j^{th}$  utterances (or any other minimal units, depending on the segmentation task considered) in the stream  $x$  and  $N$  represents the number of utterances. Different  $k$  values can be set, but it is standard to define it as:

$$k = \frac{N}{2 * \text{number of segments}} \quad (2.14)$$

### 2.3.2.3 Segmentation performance

Most segmentation efforts to date have used manual transcripts of monologue data such as broadcast news. Table 2.2 summarizes the performance of various segmentation algorithms on broadcast news datasets. Segmentation of dialogue data, and in particular multi-party meetings is considerably more difficult than monologue data. Approaches developed for text or monologue show correspondingly lower accuracy on dialogue data like the AMI corpus (Carletta et al., 2006). Table 2.3 presents the performance of various segmentation algorithms on dialogue data. It shows higher accuracies for supervised methods particularly using boundary features and discriminative SVM classifier (Georgescu et al., 2007).

Among many segmentation efforts on manual transcripts, some results using ASR output have shown little reduction in segmentation accuracy (Claveau and Lefèvre, 2015; Hsueh et al., 2006; Purver et al.,

Table 2.2 *Performance of various segmentation algorithms on broadcast news data.*

	Methods	Score	
		$P_k$	WD
unsupervised	DotPlot	-	0.44 (Claveau and Lefèvre, 2015)
	c99	0.21 (Choi, 2000)	0.36 (Claveau and Lefèvre, 2015)
	TextSeg	0.14 (Utiyama and Isahara, 2001)	-
	TextTiling	-	0.31 (Claveau and Lefèvre, 2015)
	Watershed	-	0.22 (Claveau and Lefèvre, 2015)
supervised	Maximum entropy	0.15 (Beeferman et al., 1999)	-
	HMM	0.16 (Yamron et al., 1998)	-
	HMM	0.14	-
	including prosodic features	(Tür et al., 2001)	

2006). Topic segmentation for two-person dialogue has received less attention. Few datasets are available, hence comparing system performance is difficult. However, Arguello and Rosé (2006) experimented on two corpora of dialogues between a student and a tutor in an educational domain. Their supervised classifier ( $P_k$  ranging between 0.10 and 0.40) outperformed the lexical cohesion method of Olney and Cai (2005) ( $P_k$  ranging between 0.28 to 0.49).

### 2.3.3 Topic identification

In a broad sense, topic identification is the task of identifying the topic that is related to a segment of recorded speech. Topic identification is also commonly referred to as *text classification* or *text categorization* in the text processing research community. Indeed, re-

Table 2.3 *Performance of various segmentation algorithms on dialogue data.*

	Methods	Score	
		$P_k$	WD
unsupervised	c99 (Georgescul et al., 2006a)	0.54	0.69
	TextSeg (Georgescul et al., 2006a)	0.40	0.49
	TextTiling (Georgescul et al., 2006a)	0.38	0.40
supervised	Decision tree (Galley et al., 2003)	0.23	0.25
	SVM (using boundary features) (Georgescul et al., 2007)	0.21	-

search and development in topic detection has been conducted in this community for many years. Research into text classification resulted in the production of a wide variety of practical systems, such as e-mail sorting and spam filtering, sentiment classification on customer service survey (Androutsopoulos et al., 2000; Fukuhara et al., 2007; Gamon et al., 2005, among others). An overview of common text-based topic identification techniques can be found in a survey paper by Sebastiani (2002).

The successes in text classification tasks led to a widespread adoption of the text processing techniques for speech-based topic identification. For example, Rose et al. (1991) conducted one of the earliest studies into speech-based topic identification on descriptive speech monologues with six different topics. Topic identification is commonly used in a variety of tasks to allow easier sorting, characterizing, filtering, searching and retrieving of speech data. For example, Gorin et al. (1996) used topic identification techniques in a customer service system for determining a customer's purpose of call and to route each call to an appropriate operator or automated system. Similarly, topic

detection has been employed in a banking services call center (Kuo and Lee, 2003), and in an IT service center (Tang et al., 2003).

In the domain of broadcast news (Allan et al., 1998; Doddington, 1998), topic detection would allow users to quickly locate particular stories about topics of their interest. A collection of technical papers in the area of topic detection and tracking, mainly in the broadcast news domain, can be found in a book by Allan (2012).

A variety of constraints apply to topic identification tasks. Similar to all machine learning tasks, the number of topic classes, the amount of training data available for learning a model, processing costs are some of the fundamental parameters affecting the performance of the system. Topic identification accuracy can increase as the length of the test sample (e.g. speech segment) increases. In addition to these standard constraints, Figure 2.8 provides a graphical representation of three primary constraints describing a topic identification task. Each dimension in the figure represents a specific constraint: prepared versus extemporaneous, limited versus unlimited domain, and text versus speech. The constraints on the tasks are loosened as moving away from the origin until reaching the least constrained (and presumably the most difficult) task of topic identification for an open domain, human/human conversations at the upper-back-right of the figure.

Standard classification error rate measure such as, the *recall*, *precision* and *F1-measure* metrics have often been applied to the topic identification problem in situations where single-label categorization is being applied (Hazen, 2011). The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at one and worst at zero. The F1 score is the

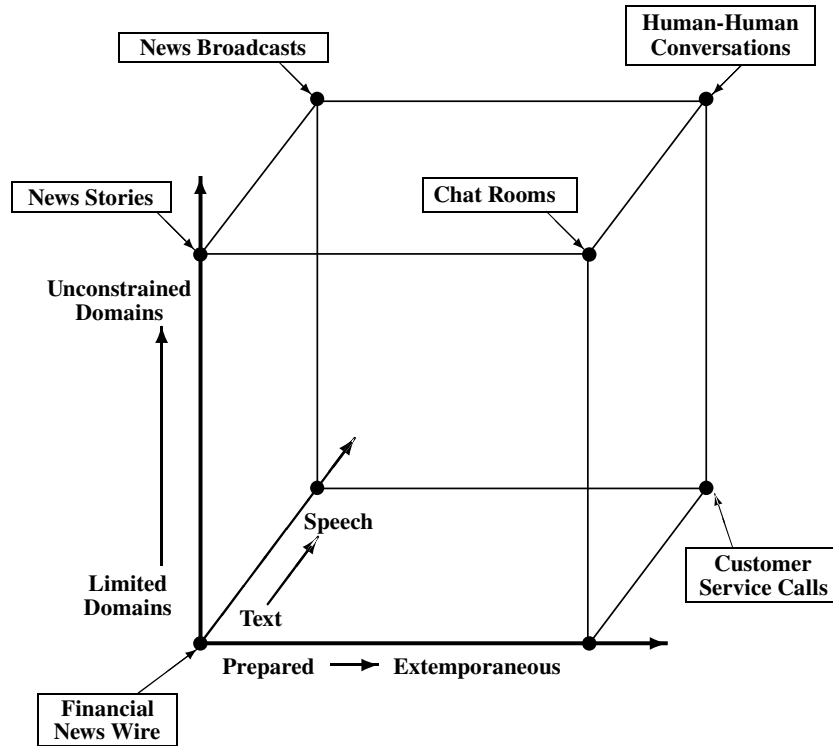


Figure 2.8 Graphical representation of three primary constraints describing a topic identification task, with example tasks for various combinations of these constraints (Hazen, 2011) (reproduced with permission from Wiley Books).

harmonic mean of precision and recall which is defined as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (2.15)$$

In detection tasks in which the goal is to detect which topics are present in a document (rather than a single-label classification), the *precision/recall curve* or the *Receiver/Operating Characteristic* (ROC) curve (Fawcett, 2006) are widely used for characterizing the relationship between misses and false alarms. The information in an ROC curve is often presented by a single scalar value known as the *Area Under an ROC Curve* (AUC) measure. AUC is the total area under the ROC curve for all false alarm rates between zero and one.

### 2.3.3.1 Technical approaches

Text classification and information retrieval techniques have been successfully ported from text processing to speech processing. Four basic steps in a typical speech-based topic identification system for converting audio documents into topic hypotheses are visualised as a block diagram in Figure 2.9.

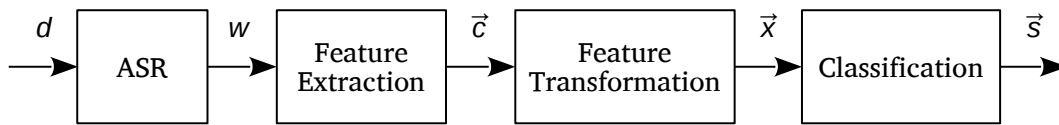


Figure 2.9 *Block diagram of the four steps typically taken by a speech-based topic identification system during the process for converting an audio document into topic hypotheses (Hazen, 2011) (reproduced with permission from Wiley Books).*

Speech-based topic identification is different from text-based topic identification as the words spoken in an audio document ( $d$ ) are not known. The audio document is first processed by an automatic transcription system hypothesizing the spoken words ( $W$ ) from the audio. The Fiscus and Doddington (2002) study on the impact of recognition errors on the overall performance of topic identification systems show that this impact is very limited compared with those obtained on clean text corpora. They found that even inaccurate transcription hypotheses could still be processed with standard text-processing techniques.

#### Feature extraction:

The most common techniques for extracting a feature vector,  $\vec{c}$ , in topic identification is the *bag-of-words* approach. The bag-of-words feature of a text (such as a sentence, an utterance transcript or a



document) is the individual counts reflecting how often each vocabulary item appears in the text. Using unigram counts, the grammar (and even word order) are disregarded. However, it is also possible to provide a richer, though higher dimensional, representation of a document by employing word n-gram counts. While the counts for the words in the vocabulary are often extracted from a single-best ASR hypothesis, they can also be estimated based on the posterior probabilities of the words present in the word lattice output of the ASR system. For instance, Hakkani-Tür et al. (2006) showed that using ASR word lattices instead of a one-best hypothesis yielded performance improvements in call classification tasks of approximately 5% to 10% relative reduction in error rate. The commonly used bag of n-gram counts is used in this thesis for extracting features from automatic transcripts of speech.

**Feature transformation:**

The bag-of-words feature vector is typically high in dimension, dominated by more frequent words such as function words, which often contain limited or no discriminative value for topic identification. This raw representation of term frequency suffers from a critical problem that all terms are considered equally important. Thus, techniques for boosting the contribution of the important content, dimensionality reduction and/or feature space transformation are commonly applied to transform the feature vectors into a feature space,  $\vec{x}$ , with a lower dimension. A commonly used technique applies weights to features based on their relative importance to the topic identification process, e.g. *Inverse Document Frequency* (IDF) weighting (Jones, 1972). In such weighting schemes, the premise is that words or terms which

occur in many documents in a collection comprising of diverse topics carry little information about a particular topic. Therefore, those terms should be de-emphasized in the topic identification process. Likewise, words that occur in only a limited subset of documents are more topic-indicative and their contribution should be boosted. The IDF weight for a term  $t$  is defined as:

$$\text{idf}_t = \log \frac{N}{\text{df}_t} \quad (2.16)$$

where  $N$  is the total number of documents in a collection and  $\text{df}_t$  is the total number of those documents that contain the term  $t$ . Using the inverse document frequency, the estimated counts of the individual features or terms of a document ( $\text{tf}_{t,d}$ ) can be represented in Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme as:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \quad (2.17)$$

For example in the task of location identification presented in Chapter 4, Appendix A presents a list of first 20 trigrams received highest TF-IDF on a document of transition-related utterances and 13 documents of room-specific utterances.

It explains for example phrases such as "*okay so*", "*am in*", "*okay so in*", "*okay i'am in*", "*into the*" are some of the most informative phrases used by most speakers as an indication of leaving a location and/or entering a new one.

The TF-IDF was originally developed in information retrieval tasks (Manning et al., 2009, chapter 6, p. 117–133) and often used in conjunction with a *cosine* distance measure to compare the similarity of two documents in a variety of application domains including speech-based topic identification tasks.

Any of the standard text document representations such as bag-of-words and TF-IDF can be very high dimensional and sparse. An alternative to the direct description of a document in a high-dimensioned term feature space is to employ latent variable modelling techniques, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), Probabilistic Latent Semantic Indexing PLSI (Hofmann, 1999) and *Latent Dirichlet Allocation* (LDA)<sup>6</sup> (Blei et al., 2003). The main premise behind these representations is that the semantic information of a document can be represented in a low-dimension space as weights over a mixture of latent semantic concepts.

The LDA model, in particular, has received a widespread recognition in the text processing and information retrieval communities for topic modelling of text corpora. One advantage of LDA is that it requires less supervision. While PLSI requires a segmented training corpus to provide direct estimates of the probability distributions over topics  $p(z)$  and documents  $p(o|z)$ , LDA takes a fully Bayesian approach. It assumes a range of possible distributions, constrained by being drawn from Dirichlet distributions. This allows a latent topic model to be learnt entirely unsupervised, allowing the model to be maximally relevant to the data being segmented and less dependent on the domain of the training set and the problems associated with human segmentation annotation.

The LDA is an unsupervised probabilistic generative model for collections of discrete data such as text corpora. The LDA models are learnt from a training corpus in an unsupervised, data-driven manner. In LDA, each document can be viewed as a mixture of a finite set of topics. In a formal definition of LDA on textual data (Blei et al., 2003),

---

<sup>6</sup>Note that the acronym LDA is also commonly used for linear discriminant analysis. Despite the shared acronym, these two techniques are not related. In this thesis, LDA refers specifically to the latent Dirichlet allocation technique.

a word is the basic unit of discrete data. Each word can be represented using a  $V$ -dimensional binary vector given a vocabulary of size  $V$ . A document is a sequence of  $N$  words denoted by  $\mathbf{w} = \{w_1, w_2, \dots, w_N\}$ , where  $w_n$  is the  $n^{th}$  word in the sequence. A corpus is a collection of  $M$  documents denoted by  $D = \{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\}$ . The LDA assumes that each document is generated using the following generative process:

1. For each document  $\mathbf{w}_m$ , draw a topic weight vector  $\theta_m$  ( $K$ -dimensional) from the Dirichlet distribution with scaling parameter  $\alpha$  :  $p(\theta_m|\alpha) = Dir(\alpha)$
2. For each word  $w_n$ , in document  $\mathbf{w}_m$ 
  - (a) Choose a topic  $z_n \in \{1 \dots K\}$  from the multinomial distribution  $p(z_n=k|\theta_m)$
  - (b) Given the topic  $z_n$ , choose a word  $w_n$  from  $p(w_n|z_n, \beta)$ , where  $\beta$  is a matrix of size  $V \times K$  and  $\beta_{ij} = p(w_n=i|z_n=j, \beta)$

Several other simplifying assumptions are made in this basic model including, the bag-of-words property of the documents and the known fixed dimensionality of the Dirichlet distribution  $K$  (and thus the dimensionality of the topic variable  $z$ ). Figure 2.10 visualises a graphical representation of the LDA model which is a three-level hierarchical Bayesian model. The only observed variables in this model are words  $w_n$  and the rest are all latent (also called hidden), which are shown by white circles.  $N$  is the number of words in the document and  $M$  is the number of documents to analyse.  $\alpha$  and  $\beta$  are dataset-level parameters representing the Dirichlet prior on the per-document topic distributions and the per-topic word distribution respectively. The variable  $\theta_m$  is document-level variable representing topic distribution for document  $m$ , and the variables  $z_n$  and  $w_n$  are word-level multino-

mial variables representing the topic assignment for  $w_n$  and the  $n^{th}$

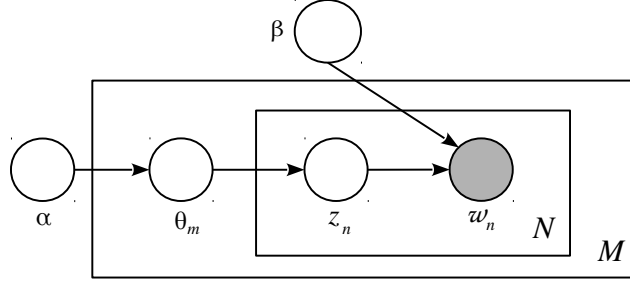


Figure 2.10 *Graphical model representation of LDA. The boxes are plates representing replicates. The outer plate represents  $M$  documents, while the inner plate represents the  $N$  repeated choice of topics ( $z$ ) and words ( $w$ ) within a document (Blei et al., 2003) (reproduced with permission from JMLR).*

word in the  $m^{th}$  document respectively. The generative process of LDA is described as the following joint distribution:

$$p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (2.18)$$

In order to use LDA, the key inferential problem is computing the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)} \quad (2.19)$$

Computing  $p(w | \alpha, \beta)$  for this distribution requires some intractable integrals. Blei et al. (2003) have shown using variational approximation work reasonably well in various applications. The approximated posterior distribution is:

$$p(\theta, \mathbf{z} | \gamma, \theta) = q(\theta | \gamma) \prod_{n=1}^N q(z_n | \phi_n) \quad (2.20)$$

where  $\gamma$  is the Dirichlet parameter that determines  $\theta$  and  $\phi$  is the parameter for the multinomial that generates the topics. The training process tries to minimise the Kullback–Leiber divergence (KLD)

(Kullback and Leibler, 1951) between the real and the approximated joint probabilities (i.e. equations 2.19 and 2.21) (Blei et al., 2003):

$$\underset{\gamma, \phi}{\operatorname{argmin}} KLD(q(\theta, \mathbf{z}|\gamma, \phi) || p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)) \quad (2.21)$$

Other training methods (Griffiths and Steyvers, 2004) use Gibbs sampling algorithm (Griffiths and Steyvers, 2002) which is based on the Markov chain Monte Carlo method:

- Go through each document, and randomly assign each word in the document to one of the  $K$  topics. Topic assignments are temporary as they will be updated in the next step. This random assignment can provide both topic representations of all the documents and word distributions of all the topics, albeit not very good ones. Temporary topics are assigned to each word in a semi-random manner according to a Dirichlet distribution. This means that if a word appears twice, each word may be assigned to different topics.
- To improve on these assignments, for each document  $d$  do:
  - Go through each word  $w$  in  $d$ :
    - For each topic  $z$ , compute: 1)  $p(z|d)$  which is the proportion of words in document  $d$  that are currently assigned to topic  $z$ , and 2)  $p(w|z)$  which is the proportion of assignments to topic  $z$  over all documents that come from this word  $w$ . Reassign  $w$  a new topic, where we choose topic  $z$  with probability  $p(z|d) \times p(w|z)$  (according to the generative model, this is essentially the probability that topic  $z$  generated word  $w$ ). In this step, it is assumed that all topic assignments except for the current

word in question are correct, and then updating the assignment of the current word using our model of how documents are generated.

- After repeating the previous step a large number of times, the assignments will eventually reach a roughly steady state where they are pretty good.

Full details of the LDA method and its variational approximation method can be found in (Blei et al., 2003).

Appendix B presents a list of {word, probability} for the 20 most probable words in 40 topics learnt on the manual transcripts of the Switchboard telephone speech corpus (Godfrey et al., 1992). By looking at the words assigned to a single class we can see they are often semantically related to one or multiple topics. For example, words in the first class are mainly related to topics like ‘money’, ‘work’ and ‘values’, words in the third class are mainly related to ‘numbers’, ‘time’ and ‘values’ or words in the seventh class are mainly related to ‘education’:

- **Topic1** *nice, dollars, working, thousand, hundred, fifty, paid, name, several, california, miles, top, absolutely, hope, except, rest, out, single, oil, guy*
- **Topic3** *two, years, three, five, four, twenty, ago, couple, six, times, half, thirty, eight, hundred, months, eighty, percent, only, major, days*
- **Topic7** *kids, school, high, usually, know, end, month, education, schools, insurance, when, health, public, up, going, having, because, side, hear, second*

The inferred vector of latent concept mixing weights for a document can be used to represent the document for a variety of tasks including topic identification, topic clustering, and document link detection (Hazen, 2011). Wei and Croft (2006) presented one of the first large-scale evaluation of the LDA modelling on information retrieval tasks. It has also been successfully applied to process auto-

matic transcription of speech data for instance in dialogue classification task (Morchid et al., 2014a) or speech-based topic detection (Morchid et al., 2014b). The LDA transformation was found to be useful for dimension reduction of a discrete form of acoustic features in broadcast media genre and show identification (Doulaty et al., 2016), acoustic information retrieval in unstructured audio analysis (Kim et al., 2009), automatic harmonic analysis in music processing (Hu and Saul, 2009) or object categorisation and localisation in image processing (Sivic et al., 2005).

### **Classification:**

Given a feature vector  $\vec{x}$ , the final step in topic identification is to generate classification scores and decisions for each topic using a variety of classification techniques. Naive Bayes, nearest neighbour and, SVM classifiers are among the most commonly applied techniques to the topic identification problem. Hazen (2011) presents a performance comparison among several different commonly used classifiers on Fisher corpus (Cieri et al., 2004). SVMs in particular, have frequently been applied in numerous speech-based topic identification studies amongst: Haffner et al. (2003); Hazen and Richardson (2008) and Morchid et al. (2014b). The SVM works with a variety of vector weighting and normalization schemes including TF-IDF and LDA. Morchid et al. (2014b) reported that employing a multi-class SVM classifier coupled LDA-based feature vector method outperforms the classification results obtained by the classical TF-IDF approach in the task of automatic theme classification of telephone conversations. They employed *one-against-one* method (Weston and Watkins, 1998) with a linear kernel for their multi-class SVM since it has been re-



ported to yield a better testing accuracy than the one-against-rest method (Yuan et al., 2012). Similar multi-class SVM classifier is employed in Chapter 4 for classifying transcription of speech segments within a simulated search and rescue conversation.

## 2.4 Topological mapping

Automatic environment mapping<sup>7</sup> has been extensively studied in the field of mobile robotics for a variety of applications such as developing fully autonomous entities capable of performing tasks in previously unknown environments, or trying to complement and help human operatives in hostile conditions of search and rescue by using rescue robots (Davids, 2002) for exploring the incident scene.

In general, the representation of a physical environment can be classified into *metric-based* and *topological-based* maps. While metric maps capture the geometric properties of the environment, the topological maps describe the connectivity of different places. Thrun (2002) presents a comprehensive introduction to the field of robotic mapping with a focus on indoor metric maps and Boal et al. (2014) provides an overview of the most prominent techniques that have been applied to topological mapping. Other types and variations of mapping techniques have been studied in the past which exploit the two basic forms of metric and topological maps either individually or in combination. Boal et al. (2014) describe a fine-grained classification for maps based on an increasing level of abstraction, comprising of metric, hybrid, topological, and semantic maps which is visualised in Figure 2.11.

---

<sup>7</sup>**Note:** This thesis introduces a speech-based approach for topological map estimation in the search and rescue context (Chapter 4.2) that is inspired, in part, by automatic topological map making algorithms that are therefore briefly reviewed here.

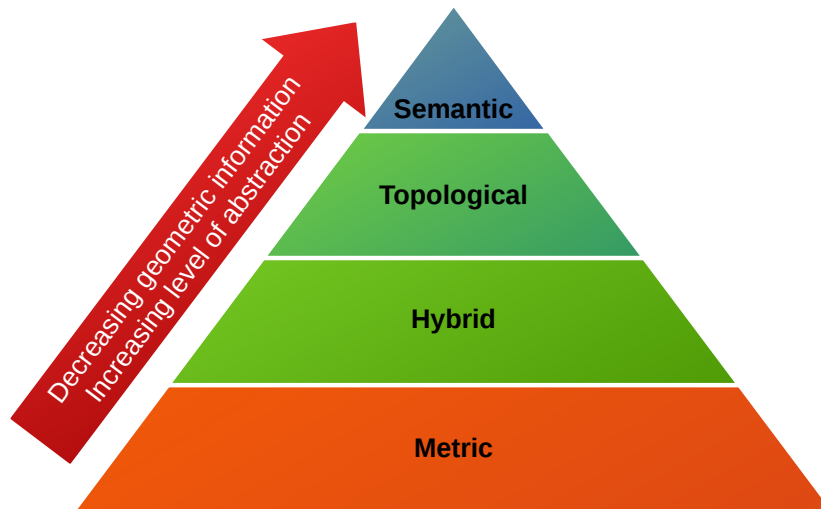


Figure 2.11 *Level of abstraction hierarchy for maps (Boal et al., 2014) (reproduced with permission from Cambridge university press).*

### 2.4.1 Topological maps

Topological maps have been mainly studied based on cognitive theories of space (Piaget and Inhelder, 1956; Lynch, 1960; Siegel and White, 1975) and mobile-robot mapping (Kuipers and Byun, 1991; Boal et al., 2014). The standard definition of a topological map (Kuipers, 1978) describes it as a graph whose vertices or *nodes* represent distinctively recognizable places in the environment, also known as *landmarks*, and the edges or *links* indicate travel paths connecting the nodes. The nodes and edges may also be annotated with higher-level semantic knowledge such as descriptions of a place, objects and semantic labelling about the environment, local and global coordinate systems, distance, direction, and procedural information on navigating between places.

While metric maps are more accurate, pure topological maps are commonly used for navigational purposes. The London Underground

network map<sup>8</sup> is a good practical example of a topological map, which presents a large-scale and complex spatial structure in an abstract form. The topological representation resembles the environmental perception and interpretation of human beings (Lynch, 1960). This is because we do not need to know where we are in millimetres and degrees to be able to navigate ourselves through an environment (Brooks, 1990). Instead, we localise and navigate ourselves using high-level information (mainly visual) about the appearance of scenes and landmarks associated with an internal representation (map) (Stankiewicz et al., 2006; Garsoffky et al., 2009).

#### 2.4.2 Automatic topological mapping

Figure 2.12 illustrates the basic steps that are typically taken by automatic topological map making techniques in the field of mobile robotics. The first step is to choose the appropriate technologies to sense the environment while an agent (often a robot) explores an area. The system acquires sensory information from one or several sources and selected features are extracted. Sensors such as range-finders and/or vision-based methods have been employed to interpret environment characteristics (see Boal et al. (2014) for a comprehensive literature on sensors and their corresponding features extraction techniques).

The next step is to detect when a new node (landmark) should be added to the map. In the process of node insertion, as the agent explores the environment, a sequence of nodes is generated in which each can be described based on the extracted features. Some of the approaches place nodes periodically either in space (displacement) or

---

<sup>8</sup>The London Underground network map (June 2016): <http://content.tfl.gov.uk/standard-tube-map.pdf>

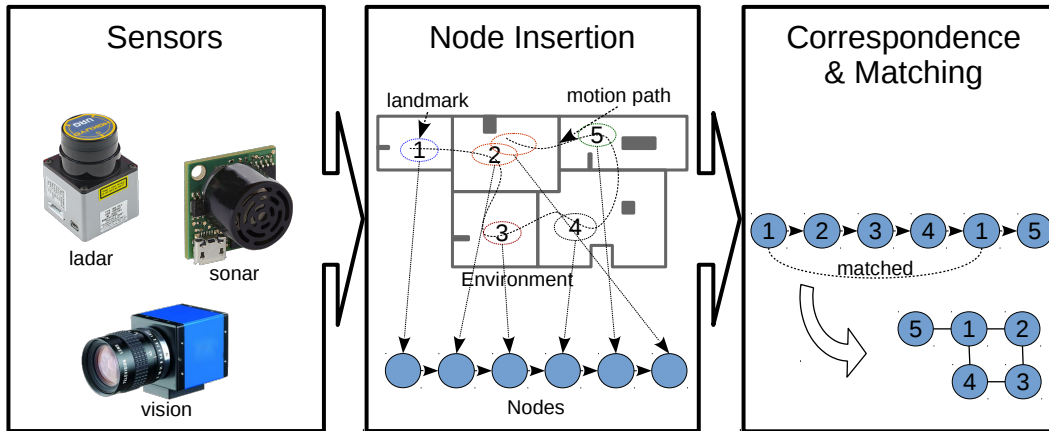


Figure 2.12 The basic steps that are typically taken by automatic topological map making techniques in the field of mobile robotics. The first step is to choose the appropriate technologies to sense the environment while a robot explores an area. The next step is to detect when a new node (landmark) should be added to the map. The final step is to determine whether each added node is a new one, or one that has been visited previously.

in time intervals. In some other strategies, such as in Tapus and Siegwart (2005), a new node is introduced whenever an important change is detected in the environment indicating that the agent has moved to a new location. This form of node insertion can produce landmarks which each represent a place that is locally distinguishable. A landmark detection approach for node insertion can produce a compact topology in which the nodes can represent a higher level of semantic knowledge.

At this stage, building a topological map is reduced to determining whether each node in this sequence is a new one, or one that has been visited previously. This involves matching the recently added node to previously detected ones, also known as the *correspondence* problem or loop-closing in topological mapping. Solving the correspondence problem is made difficult due to *perceptual aliasing* in the environment

in which different places may have a similar appearance or they may look similar to the system. In contrast, due to *perceptual variability*, a single place visited more than once can nevertheless appear distinct to the system. These problems may occur because of measurement noise, changes in the environment or illumination effects, and, in addition, a different viewpoint of the agent when revisiting a location. Failure to assess the correspondence between landmarks, increases the ambiguity of the topological map (Remolina and Kuipers, 2004). For these reasons, correspondence detection is often carried out by means of similarity distances measurements, such as the Euclidean (Goedemé and Van Gool, 2008) or cosine distances (Angeli et al., 2008).

Although several approaches have been introduced for choosing the right matches for each node and deciding on the best topological hypothesis, a robust way of dealing with unknown correspondences is to delay decision making and maintain the raw similarity scores (Ranganathan and Dellaert, 2011). This is frequently used by researchers for vision-based loop closing, in situations where the perceptual aliasing and variability are generally higher (Garcia-Fidalgo and Ortiz, 2015; Liu and Siegwart, 2013; Angeli et al., 2008, among others).



## Summary

*A brief overview of fire search and rescue voice communication system was presented at the outset. The potential role of speech technology in detection, integration, and interpretation of mission-critical information was then described. The first section of this chapter provided an outlook on the major challenges and limitations in automatic processing of voice channels in such a challenging scenario. A brief overview of the speech recognition and understanding systems that are related to the presented research work was provided in the next two sections. The final section provided a short introduction to automatic topological mapping algorithms.*

# Chapter 3

## Sheffield Search and Rescue Corpus

---

Despite the existence of language resource agencies such as LDC<sup>1</sup> and ELRA<sup>2</sup>, 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 tasks, the situation is even worse. The construction of understanding systems using statistical approaches requires suitable annotated data. In addition, due to the diverse nature of understanding tasks, datasets often need to be tailor-made to their specific needs.

This chapter is concerned with the design, corpus collection and data transcription of a new goal-oriented conversational speech corpus known as the *Sheffield Search and Rescue* (SSAR) corpus. Its design targets the task of information extraction in the context of crisis response. The first section surveys potential alternative data sources and explains the motivation behind making a new speech corpus. Then Section 3.2 presents the design of SSAR and its conver-

---

<sup>1</sup>Linguistic Data Consortium (LDC): <https://www ldc upenn edu/>

<sup>2</sup>European Language Resources Association (ELRA): <http://www elra info/>

sation task. Finally, in Section 3.5, the process of dataset collection is described by explaining the recording set-up and annotation scheme.

### **3.1 Suitable speech communication datasets**

In the application domain of extracting mission-critical information from search and rescue communications, a suitable speech dataset could be provided from three main sources: 1) radio conversation archives of crisis intervention centres; 2) available speech corpora which have been designed for similar or related tasks; 3) or making a new speech corpus by recording speech conversations during exercise sessions (either real or simulated).

#### **Using radio conversations archives:**

Technically, every search and rescue department should have their own mission radio conversations archived for later analysis. Access to these recordings is not possible without a good collaboration with a rescue department. There are several other roadblocks, such as data quality, privacy and legal limitations, which hinder us from employing them. In addition, the corresponding metadata of the conversations and their context (e.g. topic, the location of the rescue agents, their actions and information about the incident scene), which is vital for annotation, is often not recorded or provided by the departments.

#### **Using similar or related speech corpora:**

Currently available speech corpora can be used if they have the required characteristics of a search and rescue conversation for a particular task. An appropriate speech corpus for information extraction tasks should comprise meaningful conversations. Additionally,



for measuring the performance of the information extraction systems, it would be ideal if each conversation contained quantitative information about the discourse subject.

Since 1990, when the term SLU was coined by the AirTravel Information System (ATIS) project (Hemphill et al., 1990), 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 (Godfrey et al., 1992) and Fisher (Cieri et al., 2004) provide a good amount of two-party human/human conversational speech data. The original Switchboard comprises 2430 telephone conversations spoken by about 500 paid volunteers on 70 different predetermined topics. Each conversation lasts on average six minutes totalling about 240 hours of speech. The Fisher corpus includes 16454 telephone conversations averaging ten minutes in duration and totalling about 2742 hours of speech. Conversations cover 40 different topics. The Switchboard and Fisher have been extensively used in their original targeted research areas of speech recognition, speaker identification and topic detection rather than speech understanding or information extraction.

Call-Home (Canavan et al., 1997) and Call-Friend (Canavan and Zipperlen, 1996) 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 (Stein and Usabaev, 2012) (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 (Carletta et al., 2006) and DARPA-funded CALO (Pallotta et al., 2005) – 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 (Anderson et al., 1991), TRAINS (Allen and Heeman, 1995) and Monroe (Stent, 2001) corpora are collections of task-oriented dialogues. In particular, the Monroe corpus 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 dialogue behaviours. However, they are less concerned about the information content of dialogues about the discourse subject.

### **Recording exercise sessions conversations:**

In addition to regular physical training sessions, some rescue departments conduct computer-simulated exercises as well. While the real training sessions focus on the physical performance and training how to work with the equipment, the simulation training tools such as FLAME-SIM (2016) are targeting the communication performance, tactics and decision making. Although these simulated training sessions may lose some of the real characteristics of a search and rescue mission, such as high physical/psychological pressures or acoustic characteristics, the essence of teamwork and communication remains intact.

Recording these communications in a simulated environment can open up new opportunities. For example, it is possible to track rescue agents' locations, actions and the context they are in easily in a simu-

lation system. This can be valuable for providing richer and more accurate annotations. Recordings can also be performed in a controlled condition such as quiet rooms and using high-quality recording devices. Each speaker's voice and environment noise (in the simulation) can be recorded in separate channels. Such controlled conditions can result speech datasets which address particular research needs rather than dealing with different challenges such as speaker diarisation, speech recognition in noise, etc. The simulation system and the task can be designed carefully in a way to limit the conversation within the task domain while having natural and spontaneous dialogues.

Considering the data-access issues in the real mission radio conversations, and limitations on employing the available corpora in a task of situational information extraction, constructing a goal-oriented speech conversation corpus is a reasonable choice in this thesis. The next section presents a conversation task design that is used to collect speech communications in a simulated search and rescue context.

## **3.2 Conversation task design**

### **3.2.1 Conversation scenario**

Speech communications in search and rescue 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. Figure 3.1 illustrates an abstract model of the communications between first responders and task leaders using a pictographic visual language introduced by Moore (2016). In this model,

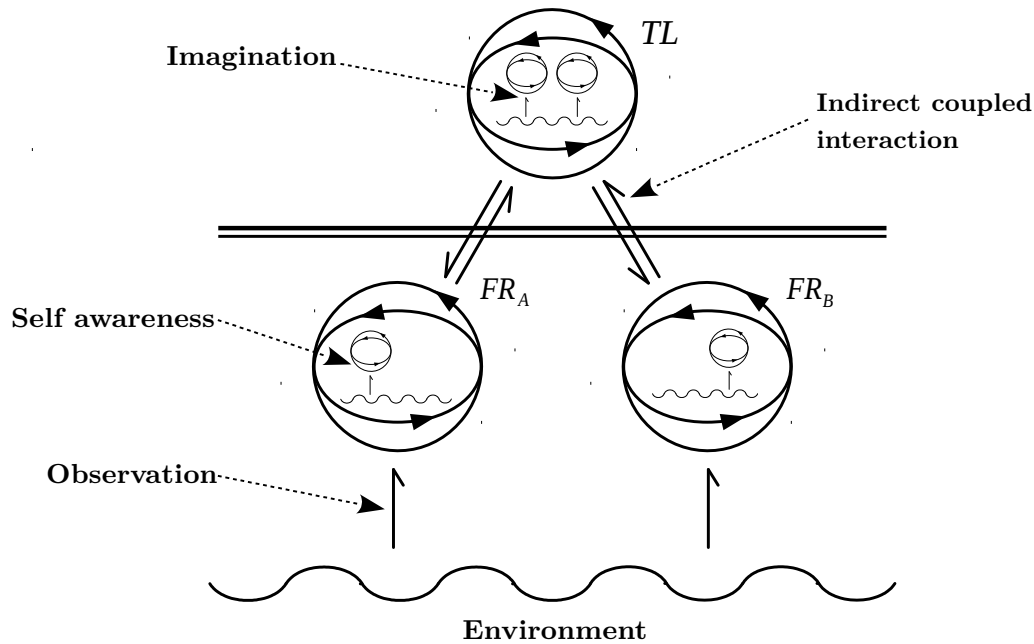


Figure 3.1 *Pictograph illustration of the abstract communication model within a search and rescue context. In this model, an individual is represented by a circle and an inner ellipse. The inner ellipse represents its thoughts and understandings. For instance here, the Task Leader (TL) has an understanding (imagination) about First Responders' (FR) status and the environment which they are in. However, each FR has an understanding about themselves and their surrounding environment (self awareness). The double arrows represent the coupled interaction between FRs and the TL that is performed remotely via a voice communication channel. In this model, FR goal is to explore the environment (i.e. incident scene) and report their observations and actions back to the control hub to update the TL knowledge about the incident scene.*

the first responders' goal is to explore the environment and report their observations back to the control hub to update the task leader'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 a first responder and a task leader. To simulate a remote conversation, they are located in separate quiet rooms. Wearing headsets, the task leader is able to hear first responder's reports and talk back for asking or

confirming any required information. The first responder is the main speaker in this task and speaks most of the times reporting to the task leader about the incident scene, their observations and actions. Given pen and paper and relying on these explanations, the task leader is asked to make an estimation of the simulated environment structure by drawing nodes to represent rooms/locations and links between them to show who they might be connected to each other. The task leader is also asked to annotate each node by writing down some of the key features about each location (e.g. room type and its condition, or key objects and their characteristics) 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.

### 3.2.2 Simulated environment design

Inspired by the simulation training systems (e.g. FLAME-SIM) which are used by some fire departments to practice their communication performance and decision making, a simulated indoor environment<sup>3</sup> was designed and built using the Unity (2016) 3D game engine. 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. Figure 3.2 shows a user-view and a top-view of one simulated environment. This figure is also overlaid with the motion trajectory of a participant and the small arrows show the viewing directions at each time.

In the SSAR task, the conversations are centred around transferring enough information about the environment from the first responder

---

<sup>3</sup>An example screen recording of a simulated environment (*Map<sub>4</sub>*) can be accessed by the following link: <https://youtu.be/X2ZAb0q35iw>

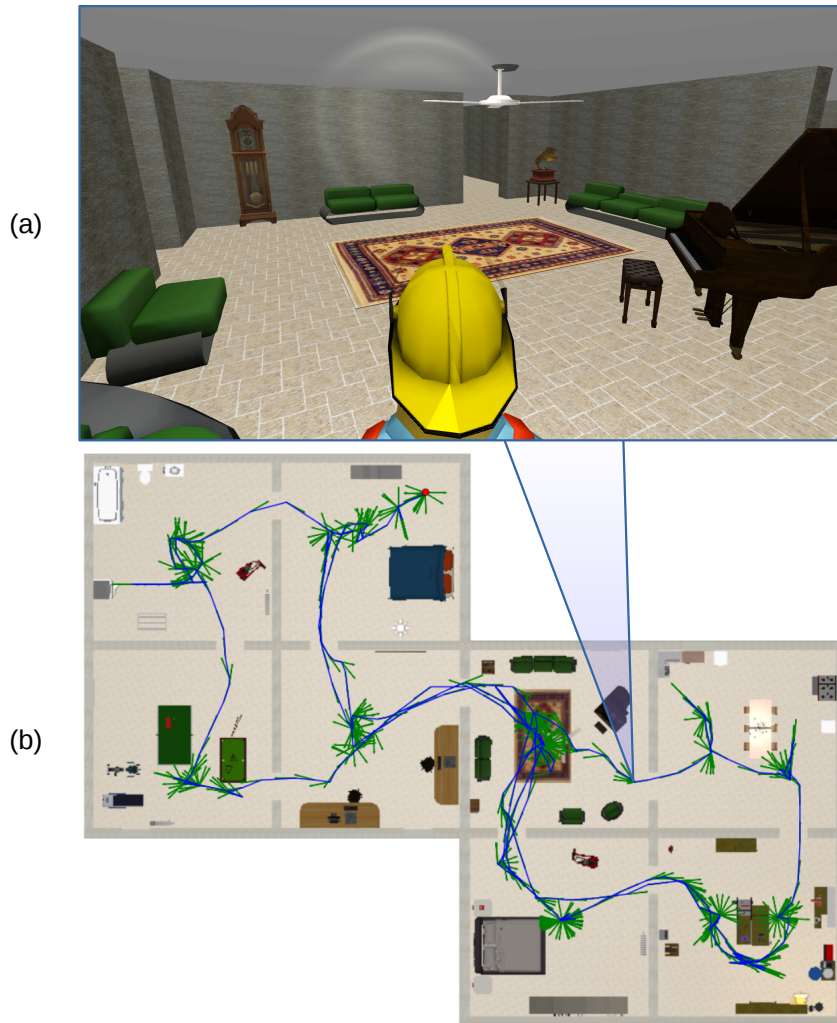


Figure 3.2 (a) A user-view of the designed simulation system. (b) A top-view of the simulated environment ( $Map_3$ ) which is overlaid with the motion trajectory of a participant and their viewing directions (small arrows) at each time.

to the task leader in order to describe the general structure. This indicates that the design of the environment map is particularly important because the more complex the structure, the more information is required to be transferred over speech channel during a successful conversation. In other words, the structural 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 in order 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 be symbolized by an undirected graph  $G = (V, E)$ . Nodes or vertices ( $v_i \in V$ ) represent rooms and links or edges between them ( $e_{ij} \in E$ ) indicate 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 (Dehmer and Emmert-Streib, 2008; Mowshowitz and Dehmer, 2012), can be used to design different map settings with a range of complexities. The topological structure of four map settings is shown in Figure 3.3. Each map setting consists of a fixed number of eight rooms. Some maps

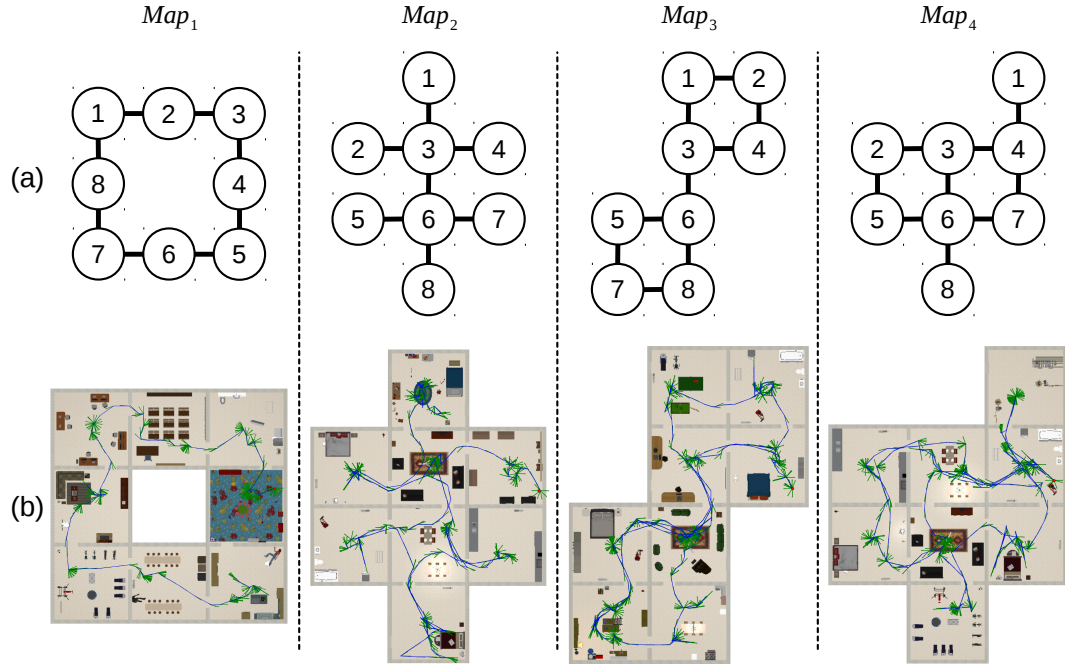


Figure 3.3 (a) The topological structure of four different map settings (*Map<sub>1-4</sub>*) which were explored by each participant. (b) corresponding top-view image of each map.

have multiple rooms of 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 thirteen different types of indoor locations (*RoomTypes*), such as *kitchen*, *bedroom* or *computer lab*, were simulated in all four map settings. Table 3.1 presents the rooms in each map and their types.

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

Table 3.1 *List of the rooms in each map and their types.*

Room	$Map_1$	$Map_2$	$Map_3$	$Map_4$
1	computer lab	kids bedroom	gym	boiler room
2	classroom	bedroom	bathroom	kitchen
3	bathroom	living room	computer lab	dining room
4	kids playroom	library	bedroom	bathroom
5	kitchen	bathroom	bedroom	bedroom
6	canteen	dining room	living room	living room
7	gym	kitchen	workshop	bedroom
8	living room	bedroom	kitchen	gym

### 3.3 Corpus recording

Recordings were performed in two separate quiet rooms for avoiding external acoustic disturbances and crosstalk between the two speakers' voice. Figure 3.4 illustrates a schematic of this set-up (top) and a picture of two participants while starting a recording session (bottom).



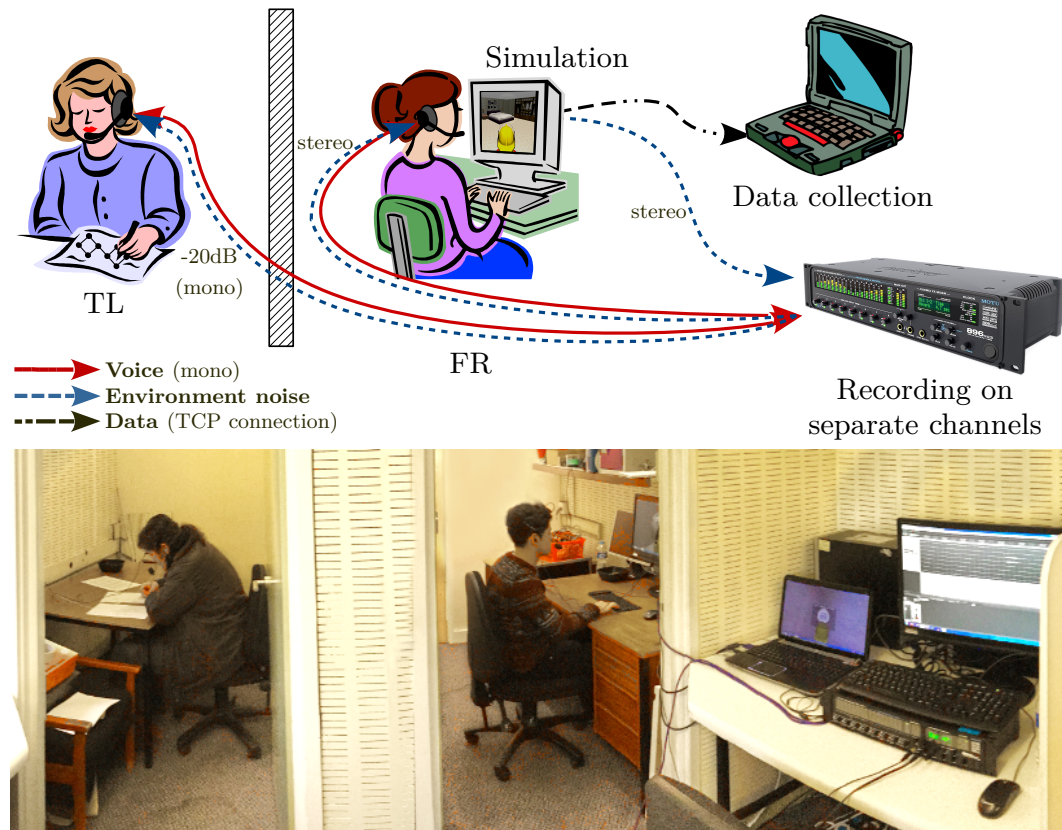


Figure 3.4 top: the recording scenario, bottom: the recording set-up in two separate quiet rooms.

Full instructions about how the task should be performed were given to the participant one day before the experiment (see Appendix C). The instruction sheet describes the recording procedure, the first responder's task of exploring and explaining, and the task leader's task of drawing an estimated topological structure of the environment based on the first responder's explanations. 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. A short practice session in a practice map was provided to each participant just before the main recording to help them familiarize themselves with how to move in the simulated environment.

The participants performed the experiment behind the closed doors by communicating with each other through the simulated remote communication system. A MOTU-896Mk3 (2016) 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 (2016)* software, was used for analog to digital 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 (*stereo*). Table 3.2 summarizes the specific recording set-up and the recording instruments used.

Table 3.2 *Recording set-up information and the specific recording instruments used.*

Microphone	Panasonic RAMSA WM-S10 (head-worn condenser)
Audio card	MOTU-896Mk3
Audio recording software	Audacity
File type	wav
Recording sample rate	48 000 Hz
Recording sample format	16-bit
Number of channels	4

Other information about the participants' motion trajectories, actions and list of objects in their field of view in the environment were logged in a computer readable text file. Figure 3.5 shows an example of such motion trajectories, an instance of a participant's field of view and surrounding objects in the simulated environment.

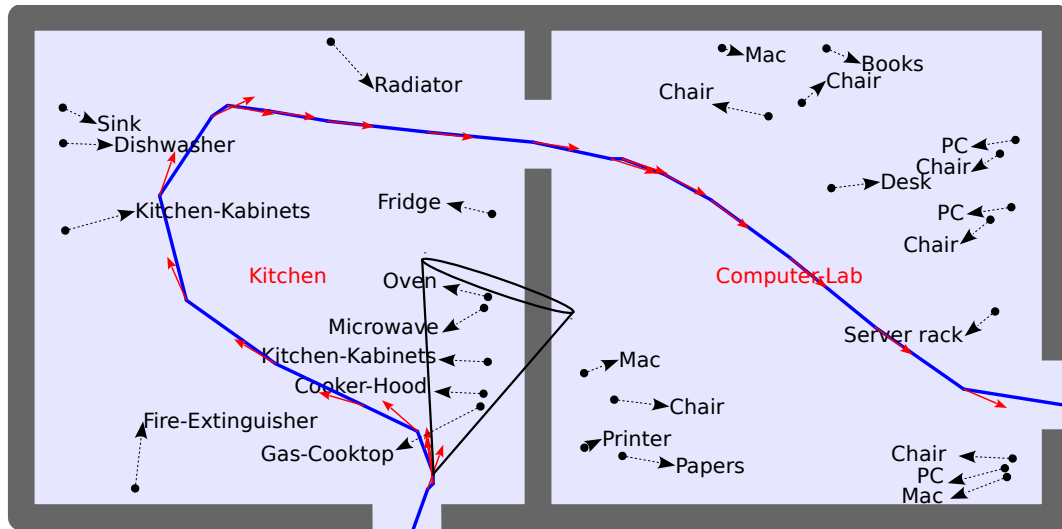


Figure 3.5 *An example of motion trajectory information plotted over the environment map, an instance of a participant's field of view and surrounding objects in the simulated environment.*

Each recording was started by the participant in the role of a first responder by pressing a *connect* button in the simulation GUI. After a successful connection, a *beep* sound was played to both participants and the same beep sound was played again at the end of each recording indicating the end of the recording. These one-second beep sounds were later used for trimming off the start and end of the recorded audio and, more importantly, for aligning the audio data with the other logged information about the participants' motion, actions and the observed objects in the environment.

A maximum time for each map was estimated based on some practice recordings during the process of the conversation task design. Maximum durations were set as six, seven, eight and eight minutes for *Map<sub>1</sub>*, *Map<sub>2</sub>*, *Map<sub>3</sub>* and *Map<sub>4</sub>* respectively. The majority of the participants (>87.5%) explored the entire area of each map in the limited time. In all experiments, the structure of the explored area of the environment was correctly estimated by the task leader. Fig-

ure 3.6 presents a hand drawing example of the  $Map_4$  estimated by one of the participants.

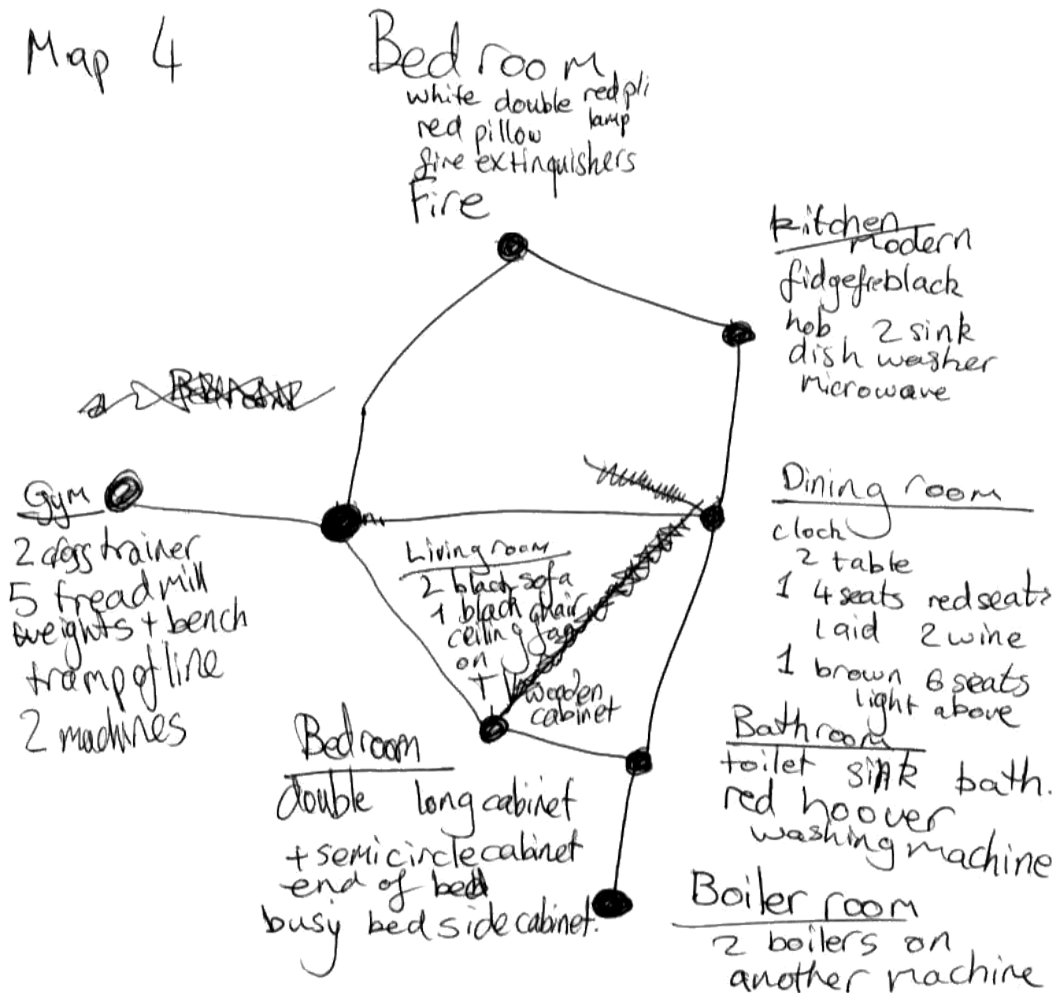


Figure 3.6 A hand drawing example of the  $Map_4$  estimated by a participant (task leader).

A total of 96 hand drawings (one for each conversation) is included in the corpus. More hand drawing examples for all four maps are available in Appendix E. Correct estimation of the visited areas confirms that the amount of information exchanged through the voice channel is sufficient for a human subject to estimate the structure of the visited parts of the environment.

### 3.4 Transcription and annotation

The start and the end of each recorded audio were first trimmed off from the beep sounds. Then in a first round, segmentation and transcription were generated automatically based on automatic transcriptions of clean speech data. The automatic speech recognition system used for the first round transcription was accessed through *WebASR* (Hain et al., 2008). Its outputs were then reformatted to XML files compatible with *Transcriber* software (Liberman et al., 1998) 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 (Moore et al., 2016). The transcription guidelines for the SSAR corpus are available in Appendix D.

Figure 3.7 presents some sections of a conversation between a first responder and a task leader 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.

### 3.5 Corpus description

The SSAR is a medium size multi-speaker corpus with 96 two-party goal-oriented spoken conversations lasting from 6 to 8 minutes each (averaging about 7.25 minutes). A total of 24 native British English speakers (66.6% Male) with a southern accent (self-reported) participated in the recording.

[ . . . ]  
**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  
[ . . . ]

Figure 3.7 *Some sections of a conversation between an FR and a TL as an example of the conversations and their transcripts in the SSAR. A ‘|’ indicates a long (about one second) pause and a ‘%’ token indicates cough/throat clearing.*

All the participants were recruited as paid volunteers through the Sheffield-student-volunteers system. The corpus totals 12 hours of speech data and about 80K words of manual transcription with about 16K vocabulary size, about 11K utterances and about 1K dialogue turns. Its perplexity against a standard *Switchboard* 3-gram language model is 173. About 11% of the utterances contain at least a token indicating aspiration, cough or throat clearing, laugh or other prominent vocal noises. Each speaker's clean speech and the environment noise are available on separate channels. This enables more control over the background noise by altering the noise level or even removing or replacing it with other noises.

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 automatic speech recognition. The current version does not include dialogue act tagging annotation.

While the SSAR has common characteristics with other corpora such as the Maptask, it has its own unique features. For instance in the MapTask, an 'Instruction Giver' can see their map and has an overview of the complete structure of the environment. However, in the SSAR, both first responders and task leaders do not have any insight into the environment structure at the beginning of a conversation. They can both discover a map gradually as the first responder explores a simulated environment. This conversation task design makes the SSAR dialogs more similar to a real Search and Rescue voice communication scenario. In contrast with the MapTask, there is no

labelling or landmarks on locations in the SSAR simulated environments. Therefore, the speakers have their own choice of vocabulary. In addition, to describe a location, one speaker can refer to a set of particular environmental features (such as objects) while another speaker may find different characteristics for describing the same location. In the SSAR four different levels of structural complexity is considered to control the information content of conversations. In the SSAR dialogues there is no eye-contact to make sure that all information about the map of the environment is transferred over the voice channel. However, in half of the Maptask dialogues, two speakers sit opposite one another and they were able to see their partner. In contrast with the MapTask conversation which there is no time limit for each conversation, in the SSAR, depending on the complexity of each map there is a limited time for exploring and discovering an environment map. The particular SSAR recording setup (separate channels for each speaker's clean speech and the environment noise) enables more control over the background noise by altering the noise level or removing or replacing it with other noises. Such recording setup provides a speech dataset addressing particular research needs for spoken language understanding rather than dealing with different challenges such as speaker diarisation, speech recognition in noise, etc. Additional information (aligned with the speech recordings) about the participants' motion trajectories, actions and list of objects in their field of view in the environment provides a form of conceptual annotation for the conversations.

The spoken conversations have many characteristics of spontaneous spoken language such as disfluencies, false starts, and colloquial pronunciations. While the dialogues are spontaneous and participants were free to talk about the simulated environment, an im-



explicit constraint is applied to these conversations by the task and the environment structure as the discourse subject.



### Summary

*To provide a suitable speech dataset in the task of information extraction from speech communication channels, a new goal-oriented conversational speech corpus was designed and collected. The SSAR corpus was recorded based on an abstract communication model between first responders and task leaders during the search process in a simulated crisis response training scenario. Each conversation is concerned with a cooperative task of exploring a simulated indoor environment by a first responder and estimating a topological map of the environment by a task leader via asking about their observations and actions. The SSAR corpus comprises of 96 dialogues between 24 speakers, totalling 12 hours, with 80K words transcribed manually. The SSAR includes different layers of annotations which can be used in a range of human/human conversation understanding tasks, automatic speech recognition and related topics. This corpus is being made available for research purposes (via LDC).*



## Locational Information Extraction

---

In the search and rescue context, spoken language is widely used for transferring critical information about the location of first responders and their ambient conditions. Automatic extraction of this information can reduce the risk of human related errors in large and fast moving operations. However, finding clear and direct references to locational information may not be possible in such highly spontaneous reports. Instead, such information may be described sporadically in multiple speaker turns or across the whole conversation (or even multiple parallel dialogues). Consequently, the information from several utterances may need to be considered for identifying locational evidence.

As highlighted in Section 2.3, the fine-grained identification of fundamental units of meaning (e.g. sentences, named entities and dialogue acts) is sensitive to high error rate in the automatic transcription of spontaneous and noisy speech. In such high error rates, there is no guarantee that it is possible to identify relevant keywords. In contrast, looking at the problem from a topic-based perspective and utilizing state-of-the-art text vectorization techniques has been

shown to result in systems that are robust to such errors (Morchid et al., 2014a;b; Hazen, 2011).

The redundancy effect has been described as the main explanation for this phenomenon (Hazen, 2011). This is due to the fact that topics are often represented by many occurrences of salient words characterizing them. When key concepts are missed or replaced (because of the automatic transcription errors), the surrounding words and phrases may help in discarding the noise and identifying the information (see Section 2.3.1). In addition, *latent concept modelling* techniques in text vectorization, such as LDA, provide the possibility of matching text segments that do not share common words. This is particularly important for being robust to the high variation in natural and spontaneous speaking style and particularly when there is limited amount of in-domain training data.

Section 4.1 introduces an approach for estimating first responder's location by framing this problem as a topic identification task on their spoken reports about their observations and actions. The location estimation is based on the notion that, significant changes in the content of a report over time may correspond to changes in the speaker's physical context. Identifying these changes can provide a rough estimation of the speaker's location. Later in Section 4.2, a similar approach is applied for *landmark detection* and *correspondence estimation* as the main steps in building a topological map (see Section 2.4) of the incident scene.

## 4.1 Speech-based location estimation

As highlighted in Section 2.1, critical information about first responders' observations, actions and events in their surroundings is often

transferred through speech communication channels. First responders' spoken reports can be viewed as verbal annotations of the incident scene. Significant shifts in the content and statistical properties of these reports would be an indication of the changes in the speaker's physical context, such as moving from one particular location to another. In addition, first responders may naturally tend to signal their task leader about their intention of moving from one particular place into another (similar topic boundaries; see Section 2.3.2). For example, words and phrases such as, "*okay so*", "*I'am in*" or "*so in this room*" might be used for signalling their intention of leaving a location or entering a new one.

Plotting a self-similarity for a conversation transcript (similar to the described DotPlotting in Section 2.3.2) can visualise such changes and transition signals. Figure 4.1 visualises an example of this self-similarity plot for a  $Map_1$  conversation in the SSAR corpus. In this example, the similarity between a pair of utterances was estimated by computing the cosine distance between their lexical frequency vectors (bag-of-words). The red dashed lines show ground-truth transition moments between rooms which were obtained from the locational annotation of the conversation. The room transition boundaries (around the red dashed lines) are showing very low similarities to the rest of the areas and high similarities to other boundaries. This can indicate that, similar words or phrases are used for signalling a location change. Observing similar patterns in most of the 96 conversations in the SSAR corpus supports the idea that identifying these signals can result in segmenting a long speech report into short units where each unit may correspond to a particular visited location. This identification can provide an estimation of the location of the speaker. The similarity between this task and the topic seg-

mentation/identification problem makes it possible to employ a wide range of techniques from the conversation topic detection field.

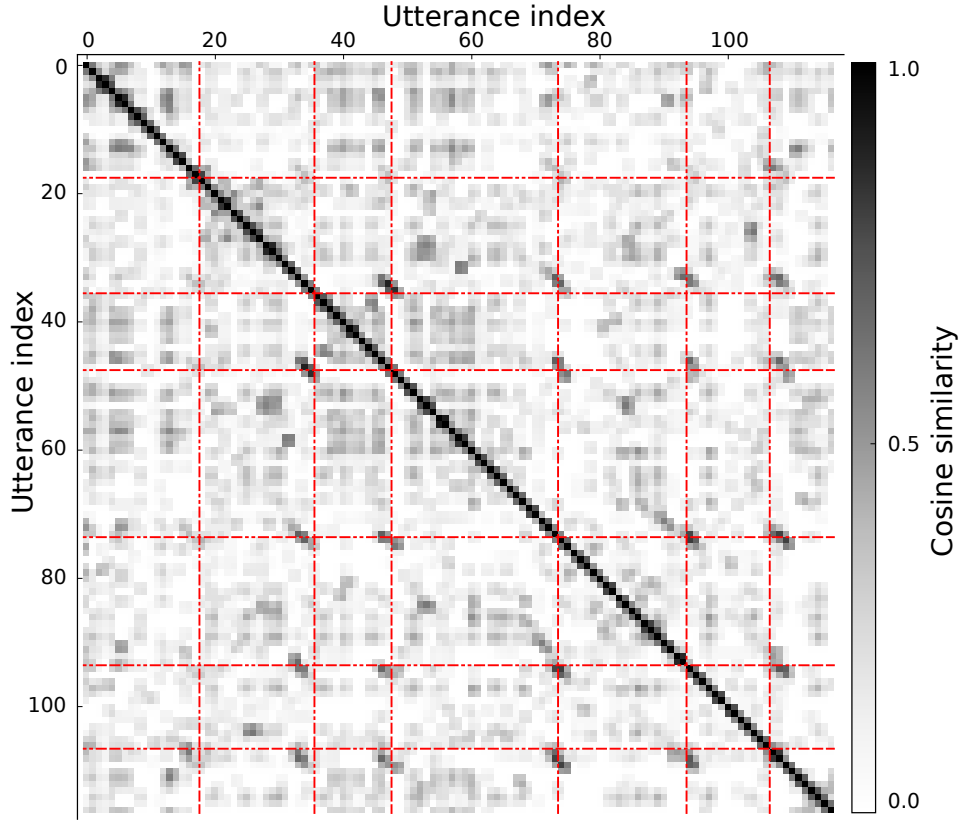


Figure 4.1 *Visualisation of self-similarity plot for one example  $Map_1$  conversation transcript in the SSAR corpus. Cosine similarity scores every pair of utterances are presented with a gray levels ranging from white for zero (no similarity), to black for one (highly similar). Red dashed lines show ground-truth transitions between rooms.*

#### 4.1.1 Transition detection

As described above, it is expected to observe distinctive boundary features when a first responder moves from one location to another. Such features can automatically be learned from labelled training data and used in discriminative approaches to identify transition times. Actual

transition times (obtained from the locational annotation of conversations) can be used to label utterances as ‘*transition-related*’ and ‘*non-transition*’ ones. To prepare the training data, a sliding window over the sequence of utterances was used. The utterances within the window was labelled as ‘*transition-related*’ if the window crossed a transition time. Otherwise, the utterances was labelled as ‘*non-transition*’.

A text document classification approach was adopted comprising three typical components of text preprocessing: document vector extraction, discriminative modelling and document classification. After a basic tokenization, a standard set of 50 English stop words were removed from the text document. Each window of utterances was presented as its raw words count vector and then projected into a vector space model based on the described LDA vectorization principle in Section 2.3.3. This represents the semantic information of a document in a low-dimension space as weights over a mixture of latent semantic concepts. A standard two-class SVM classifier with linear kernel function was trained on the collection of training vectors.

For transition detection, the same sliding window approach was applied to the sequence of automatically transcribed conversations in the test dataset. After text preprocessing and document vector extraction, the trained classifier was applied to the vector of each window for transition detection. Figure 4.2 shows a typical example of transition estimation on the automatic transcription of a conversation in the SSAR corpus.

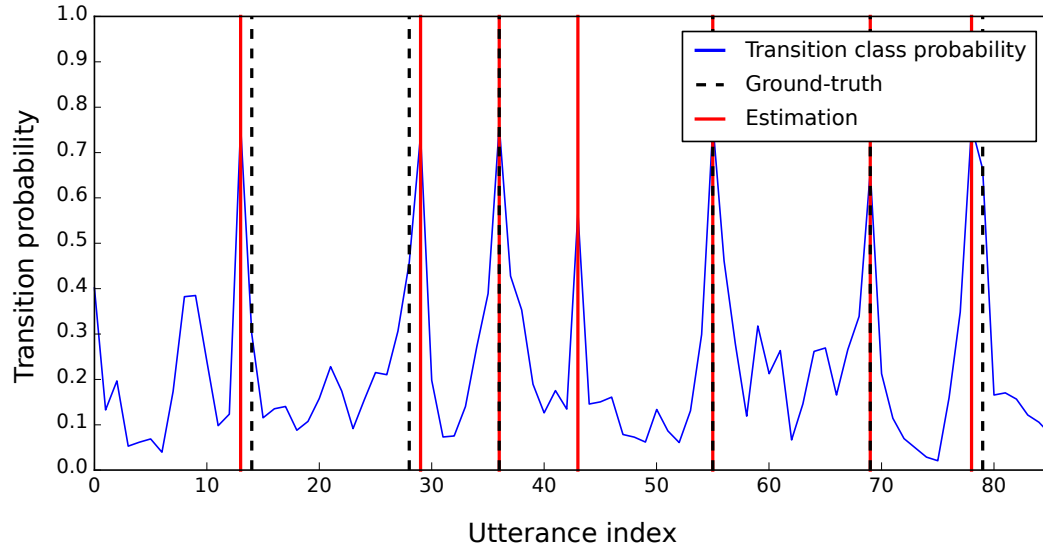


Figure 4.2 *A typical example of transition estimation on the automatic transcript of a conversation in the SSAR corpus. A sliding window with the size of three was used. The ground-truth and the estimated location transition lines are plotted. The blue line shows the transition class membership probability estimated by the SVM classifier.*

The transition class membership probability estimates are shown by the blue line. The vertical black dashed lines show ground-truth transition moments and each vertical red line indicates an utterance which is classified as transition-related. Over-segmentation can happen in the transition detection mainly in high ASR WER. An example of this over-segmentation can be found at utterance 43 in Figure 4.2. As a result of using a window of utterances, often two or, in some instances, three successive utterances were estimated as transition-related. In these cases, the one with highest transition class membership probability was selected as the transition moment.



### 4.1.2 Location estimation

Changes in the information content of a report over time would correspond to shifts in the speaker's physical context. Identifying these changes can provide a rough estimation of the location of the speaker. Recalling Section 2.3.2, topic segmentation is often the first step before topic identification. The estimated transition times by the segmentation process can be used to divide a full sequence of utterances into smaller sections where each section is likely to be related to one location. Similar to topic identification tasks, taking each segment as a whole into account can result in a more accurate identification in a high word error rate of automatic transcriptions compared with a single utterance or a short window.

To prepare the training data, all the successive utterances in a room were considered as a single training example and labelled as its corresponding *RoomType*. Low-dimensional training vectors were produced based on the described preprocessing and LDA vectorization principle in the segmentation step. A multi-class SVM was then trained on the training vectors.

For location identification, after estimating a transition time and providing a new segmentation point, the entire segment was projected into the vector space model. Given this vector and using the SVM classifier, each segment was classified as its most likely related *RoomType*. This can provide an estimate of the *RoomType* in which the first responder was at each time. Figure 4.3 illustrates a typical example of the *RoomType* probability estimation for each segment of the conversation that was used in Figure 4.2. In this example, the first segment is clearly more related to the *RoomType*1 than the rest. However,

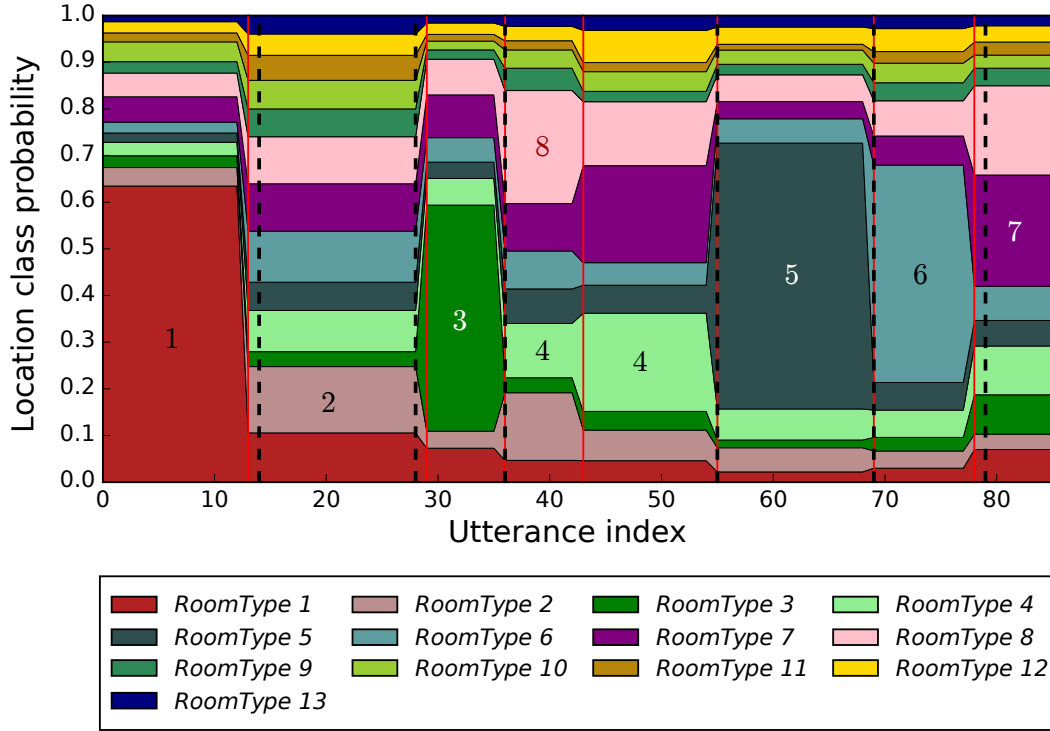


Figure 4.3 A typical example of the location identification on the automatic transcript of a conversations in the SSAR corpus. This shows the SVM class membership probability distribution for 13 RoomType classes estimated for each segment. In this example a participant visited rooms in the following order:  $R1 \rightarrow R2 \rightarrow R3 \rightarrow R4 \rightarrow R5 \rightarrow R6 \rightarrow R7$ . The estimated sequence of visited locations is:  $R1 \rightarrow R2 \rightarrow R3 \rightarrow R8 \rightarrow R4 \rightarrow R5 \rightarrow R6 \rightarrow R7$ .

an over-segmentation in *RoomType4* resulted in a short segment and consequently misclassification of the fourth segment as *RoomType8*.

#### 4.1.3 The impact of transcription errors

The proposed speech-based location estimation system was evaluated on the SSAR corpus presented in Section 3. The SSAR was divided based on a K-Fold cross-validation scheme with  $K = 12$  into a *training*, *test* and *development* (a held out tuning dataset) datasets with 80, 8 and 8 conversations respectively.  $K = 12$  was used to have at least two examples of each map in the test and development datasets. This

is important because not all *RoomTypes* exist in each map and in some conversations the entire map is not fully explored. Each of these datasets was created randomly by selecting full-length conversations from all four maps equally. This means the test and development sets each comprised a total of eight conversations (two from each of the four maps) and a total of 80 conversations for the training set (20 from each map).

To investigate the impact of transcription errors of noisy speech on the location estimation performance, a set of experiments was conducted on the automatic transcripts of the test data with different noise levels. As a difficult babble noise for ASR systems, the CHiME-3 café (CAF) noise (Barker et al., 2016) was used to reduce the automatic recognition performance. Clean speech data of the test-set was mixed with the noise, making different noise levels ranging from clean speech to a *Signal-to-Noise Ratio* (SNR) of about 0 dB with intervals of 5 dB. The original environment noise of the conversations (included in the SSAR) were not used because, the noise level changes during a conversation. The automatic recognition of these noisy data provided transcriptions with different WERs ranging from 16.1% to 96.8%.

#### 4.1.3.1 Automatic transcription system

A standard GMM/HMM ASR system was used for transcribing the test-sets. The detailed system is described in Section 5.2.1. The acoustic models used for the experiments were trained on approximately ten hours of clean speech data in the training dataset of each fold using the Kaldi open-source speech recognition toolkit (Povey et al., 2011).

For decoding a 16K lexicon, a trigram language model was made by interpolating a background and an in-domain language model. The background language model was a relatively large model trained on the transcriptions of the Switchboard telephone speech corpus (Godfrey et al., 1992) with approximately three million words. The in-domain language model was trained on the annotations of the training subset with about 65K words. Both of the models were trained and interpolated using the *SRI Language Modeling toolkit* (SRILM) (Stolcke, 2002). The interpolation weight of these two models was tuned using the independent development set.

The trained ASR system was used for transcribing the test-sets with different noise levels. The dashed line in Figure 4.5 shows its performance in terms of WER for each SNR. This system achieved a WER of 16.1% on the clean speech version of the test-set. This ASR system was intentionally trained on clean speech data only but used for decoding the noisy data of the test and development sets. As such, this recognizer has an acceptable performance on the clean data, but its WER rises significantly on the impact of noise level increase compared to state-of-the-art systems. This allowed the location estimation system to be examined under a good range of WER with a relatively low computational cost for ASR acoustic model training.

#### 4.1.3.2 Document classification system

The ASR transcripts were first preprocessed by tokenizing and removing a standard list of stop words using the *Natural Language Toolkit* (NLTK) (Bird et al., 2009). The documents vectors were then produced by applying the LDA scheme based on the Gensim topic modelling framework (Rehurek and Sojka, 2010) implementation of

the LDA for learning text topic models. The LDA models were trained on the manual transcripts of the Switchboard telephone speech corpus (Godfrey et al., 1992) in an unsupervised, data-driven manner with symmetric priors. The Switchboard was used since it is a large natural conversation corpus (2,400 conversations) with a broad range of topics (about 70) which makes it suitable for training a rich topic model.

The number of LDA topics was tuned on the transcription of the development dataset on the task of location identification. Ground-truth information about the transitions between rooms was used to divide the utterances in each conversation. LDA models with different numbers of topics, ranging from 10 to 100 topics were used to classify each segment. Figure 4.4 illustrates the systems performance (F1-score) as a function of number of LDA topics. Each experiment were performed five times with different initial LDA topics. The number

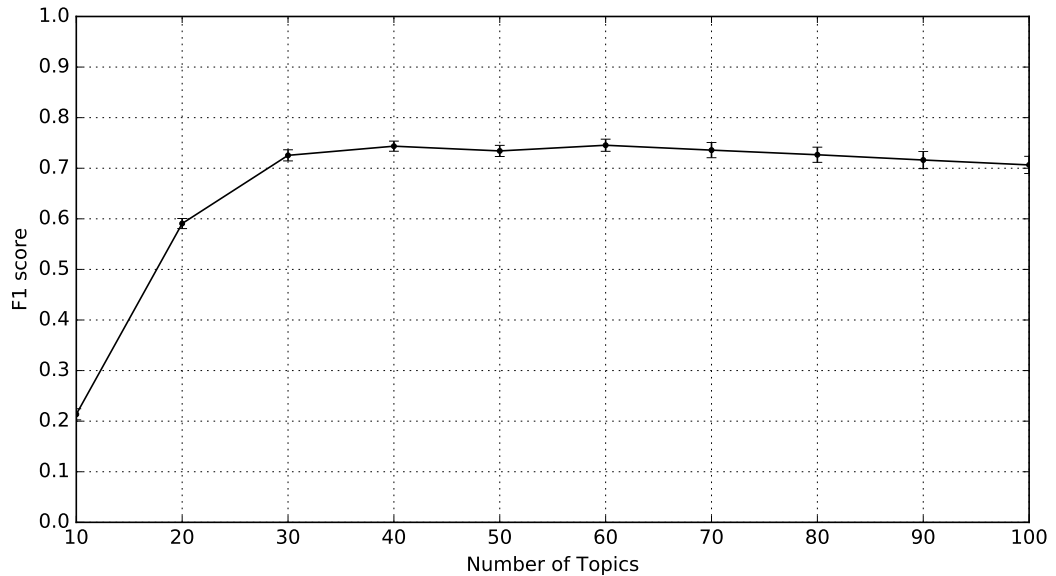


Figure 4.4 *The location identification performance on transcription of the development dataset as a function of number of LDA topics. The performance on each number of topics presents the average of five experiment with different initial LDA topics.*

of LDA topics was set to  $T=40$  which provides a high performance as well as low computation power in compare with other tested number of LDA topics.

The development dataset was also used for tuning the size of the sliding window (3 utterances) in transition detection. The LDA topic posteriors for each document were used as its vector representation. The SVM classifiers were trained on the low-dimensional LDA representations of labelled conversation segments in the training set using the *Scikit-learn* (Pedregosa et al., 2011) implementation of the SVM.

In order to investigate the effect of LDA vectorization, another set of experiments was conducted using the TF-IDF vector representation. The SVM classifiers were trained on the TF-IDF representations of labelled conversation segments in the training set to be used for transition detection and location identification.

#### 4.1.3.3 Results

**Transition detection:** The described WD-score in Section 2.3.2.2 was used as the quality measure for the transition detection. Figure 4.5 presents the results obtained by the transition detection step on the automatic transcription of test data with different SNRs. WD errors of both LDA and TF-IDF methods are shown to compare their performance. The results in this graph show a lower WD error using the LDA-based approach compared to the TF-IDF implementation. This confirms the positive effect of the LDA vectorization approach.

It is notable that on both systems, in spite of a statistically significant increase in WER of about 9% (from 16.1% WER on clean data to about 25% WER on an SNR of 20 dB), the segmentation error did not receive a high negative impact. In fact, its WD error experiences a

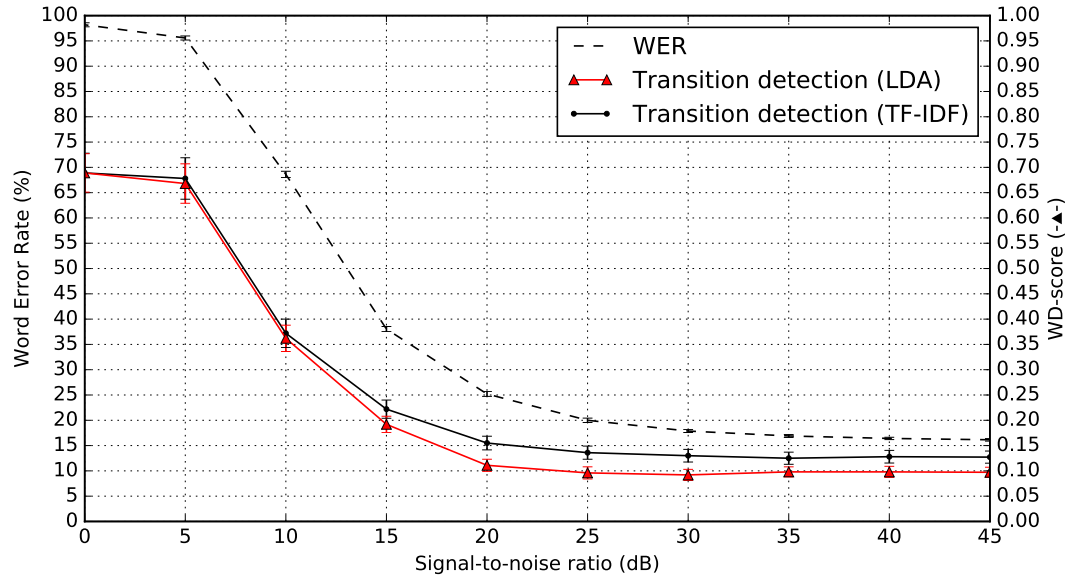


Figure 4.5 The ASR transcription WERs on different SNRs are shown with a dashed line. The red line shows the WD errors of the LDA-based method for transition detection on the automatic transcription of test data. The black line shows the system performance using the TF-IDF vector representation.

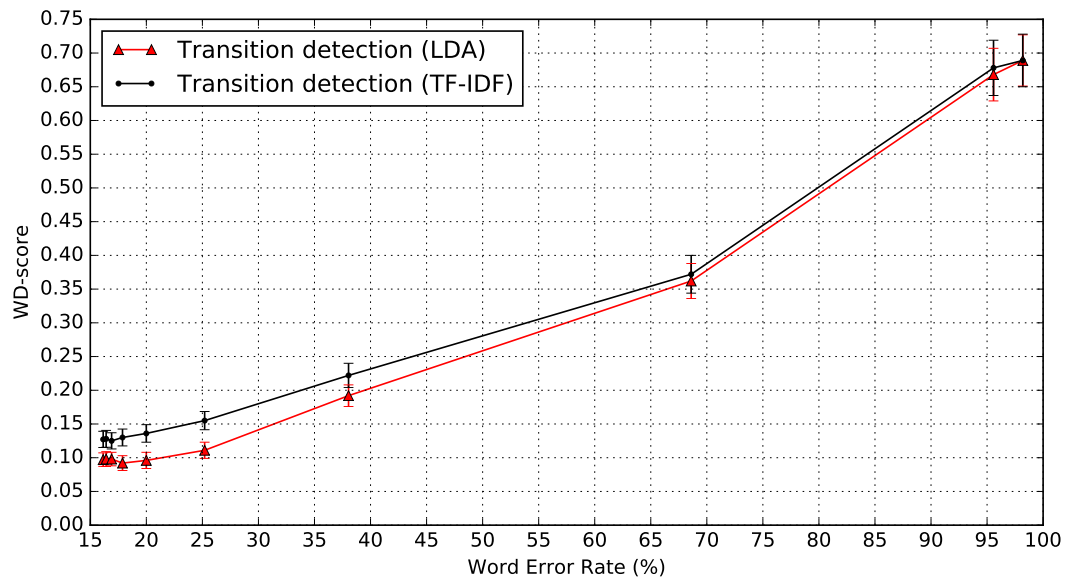


Figure 4.6 The transition detection performance on different transcription WERs. The red line shows the WD errors of the LDA-based method for transition detection on the automatic transcription of test data. The black line shows the system performance using the TF-IDF vector representation.

tiny increase of about 0.02 for the LDA-based approach, which is not statistically significant ( $p=0.09$ ). However for SNRs lower than 20 dB, the WD error increase is statistically significant. To visualise the effect of WER increase on the transition detection error, Figure 4.6 illustrates the WD scores as a function of WER. For WERs greater than about 25%, the WD shows no or very little robustness to WER increase on both systems.

**Location identification:** The F1-score, as one of the most commonly used ‘*single number*’ measures in topic identification and information retrieval tasks (see Section 2.3.3), was used to measure the location identification performance. Since F1 was calculated at the utterance level, it reflects both identification and segmentation performance together. Therefore, this performance can be considered as the overall quality of location estimation. Figure 4.7 illustrates the location identification performance obtained on the automatic transcription of test data with different SNRs. In this figure, the red line presents the LDA-based identification performance as a function of background noise. In comparison, the performance of the TF-IDF-based identification system is shown with a black line. About an average of 19% relative gain by using the LDA method shows the positive effect of the LDA vectorization on the system performance compared to the TF-IDF system at SNRs ranging from 45 dB to 10 dB.

Figure 4.8 presents the same F1 performance results as a function of WER. The results illustrate a moderate performance drop under the effect of WERs increase from 16.1% to about 70%. In this range of WERs, the location identification performance (for the LDA-based approach) experienced a decrease about 37% slower than the WER increase. This moderate performance decrease demonstrates a level of robustness to the high WER in the automatic transcription of noisy



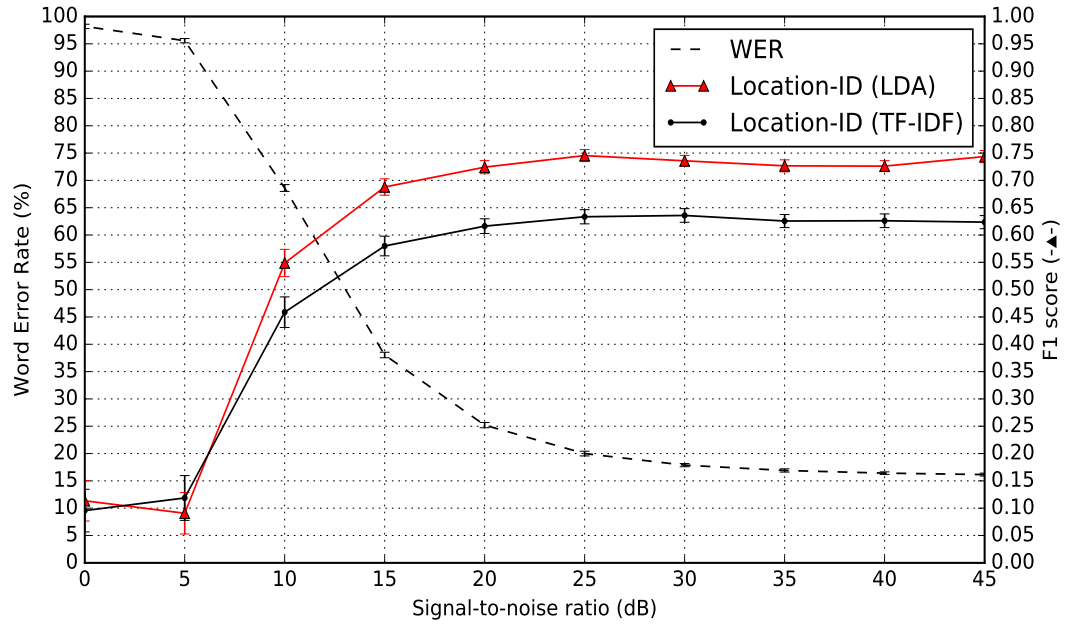


Figure 4.7 The ASR transcription WERs on different SNRs are shown with a dashed line. The red line presents the LDA-based location identification performance (F1). The black line shows the system performance using the TF-IDF vector representation.

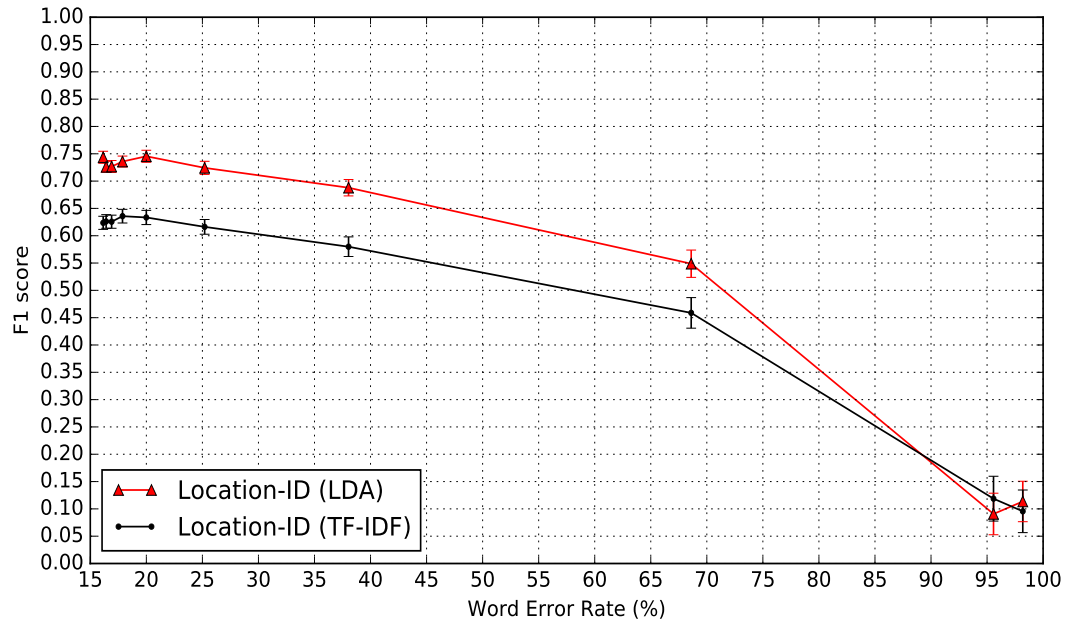


Figure 4.8 The location identification performance on different transcription WERs. The red line presents the LDA-based location identification performance (F1). The black line shows the system performance using the TF-IDF vector representation.

speech. The location identification failed on extremely inaccurate transcriptions of noisy speech. The experimental results confirm the potential application of a topic-based perspective in providing an estimation of first responders' location during a search and rescue mission based on their spoken reports.

## 4.2 Speech-based topological map estimation

An understanding of the incident scene layout is one of the main enhancing factors for situational awareness formation and the efficacy of a search and rescue response. Different strategies for automatic map making have been explored mainly within the field of mobile robotics (see Section 2.4). Techniques have been introduced for estimating metric-based or topological-based maps by interpreting the information provided by a robot (or multiple robots) while it is probing the environment using a variety of sensors.

A first responder's spoken report is a verbal description of what has been observed while exploring an incident scene. The stream of information in such a report can be compared with the robot sensory data. It can be hypothesised that similar strategies in mobile robot mapping can be adopted for estimating the environment structure from spoken reports. However, there are indisputable differences between these sources of information. In contrast with robot sensory data which is often well structured metric values, such a detailed information is very unlikely to be found in these reports. Instead, the spoken reports are generally about the main reference points such as locations, events and so forth. Consequently, estimation of a metric-map from spoken reports seems out of reach.

Unlike the detailed information provided in metric-maps, as described in Chapter 2.4, a topological map represents the structure of a physical environment as an abstract graphical model consisting of nodes and edges (Boal et al., 2014; Kuipers and Byun, 1991). These light-weight maps can represent higher-level semantic knowledge, such as how particular locations are linked to each other. Such a high-level representation methodology is more similar to the environmental perception and interpretation of human beings (Lynch, 1960), which makes it more applicable in a speech-based mapping problem.

The previous section presented how a topic-based approach can be employed for tracking the changes in the information content of each report for providing an estimation of a first responder's location during a search and rescue response. Looking from a similar perspective, a speech-based approach is introduced in this section for performing two primary steps in topological map making on the spoken reports. First, detecting when a new node should be added to the map and then, estimating the correspondence of the recently added node to the all previous ones (see Section 2.3.3).

#### **4.2.1 New node detection**

A new node can be introduced to the map whenever an important change is detected in the environment as an indication that the agent (i.e. a robot in mobile robot mapping) has moved to a new location (see Section 2.4). Detection of these location landmarks results in a compact map which the nodes can represent a high level of semantic knowledge.

The previous section showed that topic-segmentation techniques, such as identifying distinctive transition features in the spoken reports, provide an estimation of the time that a first responder has moved into a new location. The estimated transition times were used to divide a full sequence of utterances into smaller sections which each is more likely to be related to one location. These segments of the report were shown to provide enough information about the particular locations of the speaker. By this means, each segment can offer the required characteristics to be considered as a topological landmark.

The segmentation strategy described in the previous section was used for identifying transition moments and segmenting the spoken report. A node was allocated for each segment and the utterances in each segment were retained as a fingerprint of the node. A sequence of nodes was gradually formed by allocating a node for each segment of the utterances. This process is illustrated in Figure 4.9.

#### 4.2.2 Correspondence estimation

At this stage, identification of the global structure of the environment as a topological map is reduced to determining whether each node is a new one or if it has been visited previously. Here the utterances in each node are the information source for estimating the correspondence between a pair of nodes. The correspondence estimation was carried out by means of similarity measurements between each pair of nodes.

The correspondence estimation problem is made difficult due to the fact that the spoken reports in different places may appear similar because of the automatic transcription errors. This is analogous to the perceptual aliasing problem in automatic mapping (see Sec-

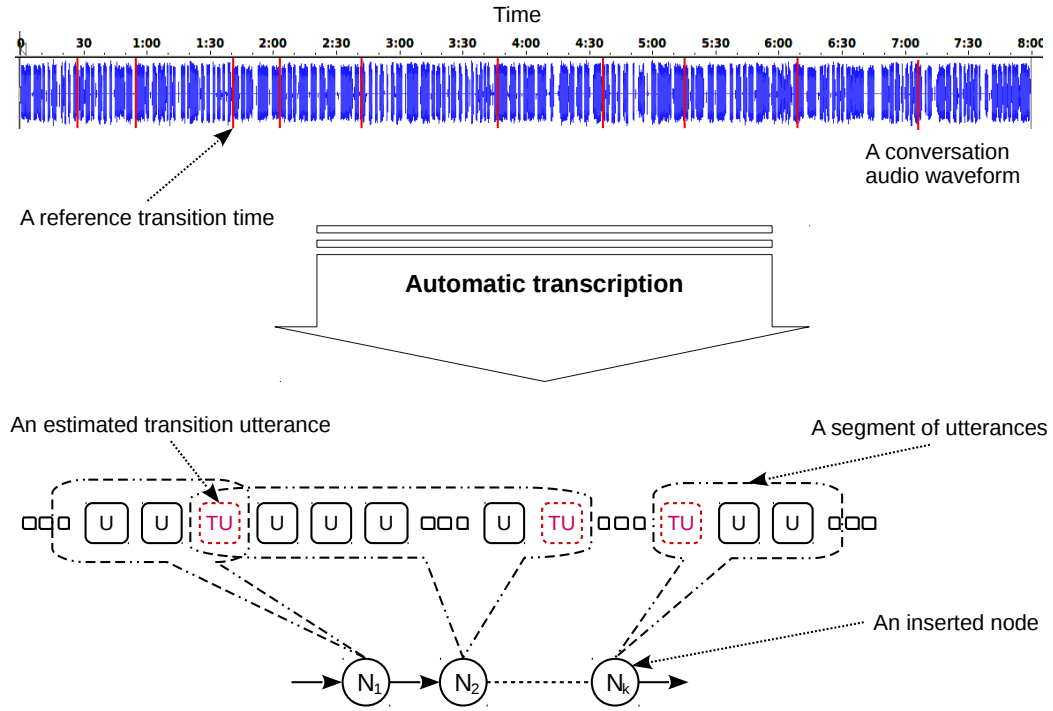


Figure 4.9 *New node insertion by identifying transitions utterances (TU) in the automatic transcription of a spoken report. Each node comprises the utterances of its corresponding (U) segment.*

tion 2.4). To reduce the perceptual aliasing posed by the ASR errors and improve the distinctiveness of nodes, the entire segment of utterances in each node was considered as a single ‘document’ to be compared against other nodes. Comparing a large segment can increase the redundancy which can lead to higher robustness to transcription errors.

The cosine distance, as an extensively used similarity measure (Han et al., 2011) in information retrieval and data mining, was used to estimate the correspondence of two nodes  $x_i$  and  $x_j$  by computing

the similarity of their document vectors  $\vec{D}_i$  and  $\vec{D}_j$  as:

$$\text{cosine}(\vec{D}_i, \vec{D}_j) = \frac{\vec{D}_i \cdot \vec{D}_j}{\|\vec{D}_i\| \|\vec{D}_j\|} = \frac{\sum^v d_i \cdot d_j}{\sqrt{\sum^v d_i^2} \sqrt{\sum^v d_j^2}} \quad (4.1)$$

where  $d_i$  and  $d_j$  are components of vectors  $\vec{D}_i$  and  $\vec{D}_j$  of size  $v$ .

In addition to the aliasing issue, it is almost impossible that two nodes are identical. This is because the same place may be described differently at different times. This problem is comparable with the perceptual variability problem in automatic mapping (see Section 2.4). In order to reduce the variability problem, each document was projected into an LDA vector space model to be able to compare the semantic information of documents independent of the vocabulary used (see Section 2.3.3). Using such a vectorization approach, two nodes can be scored with a high correspondence value as long as their documents are semantically related and even if they do not share similar vocabulary.

The correspondence between the most recent node ( $x_n$ ) with all the previously detected ones ( $x_i \forall i \in \{1, 2, \dots, n-1\}$ ) was computed as:

$$c_{n,i} = \text{cosine}(L\vec{D}A(D_n), L\vec{D}A(D_i)) \quad (4.2)$$

where  $D_n$  denotes the document of node  $x_n$  and the  $L\vec{D}A(D_n)$  is the LDA vector representation of the  $D_n$ . For a total of  $N$  nodes, all the correspondence values were presented as a  $N \times N$  symmetric matrix with a diagonal of one, as each segment matches itself. Figure 4.10a visualises a typical example of the estimated matrix  $C$  for a  $Map_4$  conversation in comparison with its ground-truth matrix  $GT$  (Figure 4.10b) which was made based on the SSAR location annotations.

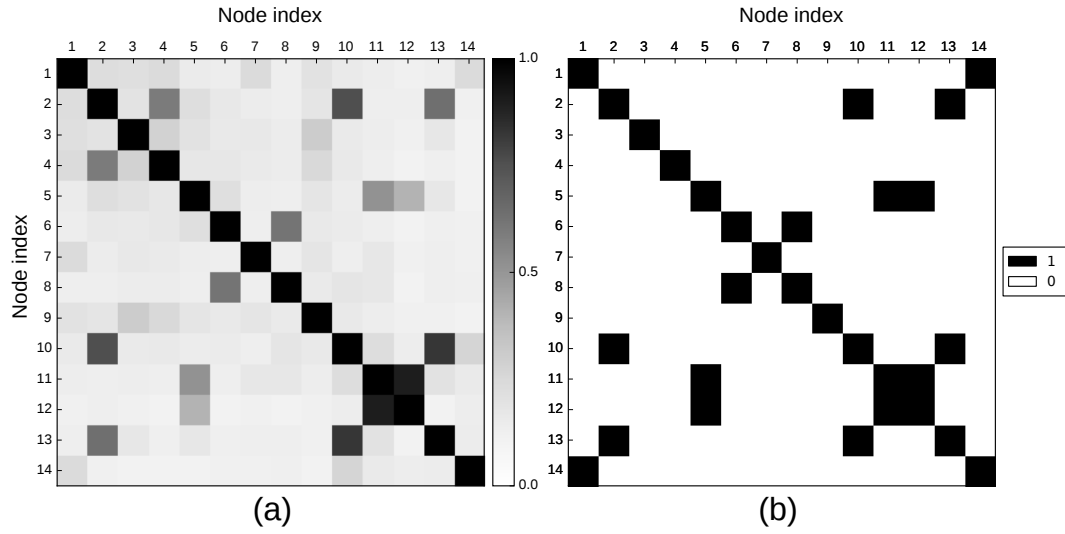


Figure 4.10 (a) Visualisation of an estimated correspondence matrix ( $C$ ) for a  $Map_4$  conversation example. Correspondence scores are presented with gray levels ranging from white for zero, as an indication of no match between a pair of nodes, to black for one as a full correspondence. (b) The ground-truth correspondence matrix ( $GT$ ) of the conversation.

The estimated correspondence scores are presented with gray levels ranging from white (*zero*), as an indication of no match between a pair of nodes, to black (*one*) as a full correspondence between them. The 11th and 12th nodes on the  $GT$  are representing a single location which shows an over-segmentation in the node insertion process. Such over-segmentations can generate short segments which may contain insufficient information to be compared with the other nodes. However, if the information content of each fragment is enough, they receive a high correspondence value.

A likely topological representation of a matrix  $C$  can be generated by thresholding its correspondence values and converting them into a binary form to be used for closing the loops and constructing the topological map graph. However, a loop-closing decision once taken cannot be reversed and this removes valuable information.

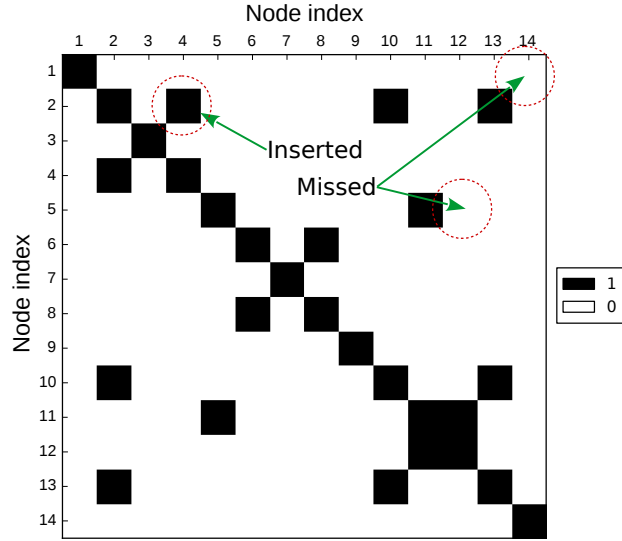


Figure 4.11 *Visualisation of the estimated correspondence matrix  $C$  presented in Figure 4.10a after converting it into its binary form by applying a threshold of 0.5.*

The estimated correspondence values are often preferred to retain without modifications for use in combination with other information scores such as a prior map or information from other mapping systems like odometry (see Section 2.4). However, for visualising a likely topological representation of the example above, a threshold of 0.5 was applied to the estimated  $C$  in the Figure 4.10a. This thresholding converted the matrix  $C$  into a binary form which is shown in Figure 4.11. In this case, this thresholding resulted in missing two correspondences and adding an incorrect one. Figure 4.12 illustrates graphically how the nodes sequence of this example can be folded into a likely topological map based on the estimated binary correspondence between each pair of nodes. The over-segmentation error at the 11th and 12th nodes is masked in this case. The correspondence between the first and the last segments is missed and, the second and the fourth segments are misidentified as a same location. These errors resulted in a



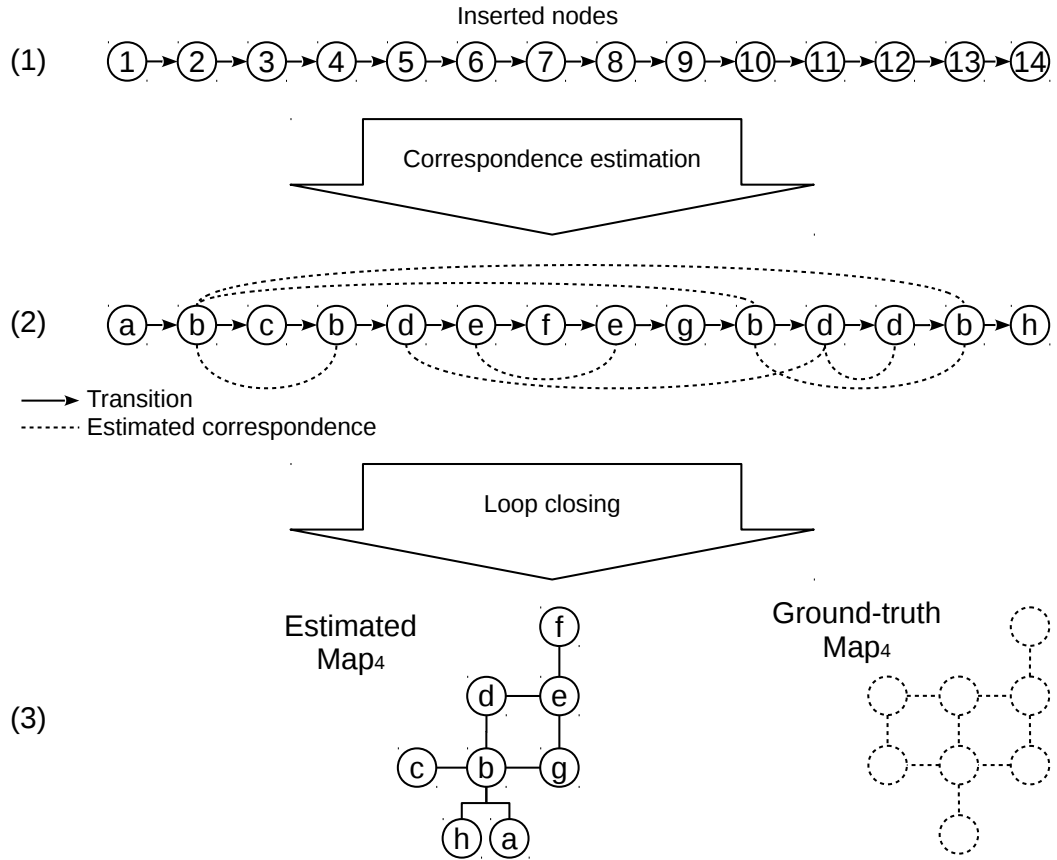


Figure 4.12 *A graphical visualisation of folding a sequence of estimated nodes from places which appear to correspond with each other and transforming it into a likely topological map.*

slightly different topological map with one edge substitution and one edge deletion.

### 4.2.3 Experiments and results

Based on the experimental set-up in location identification (see Section 4.1.3), a set of experiments was conducted on the automatic transcripts of the test data with different noise levels. A new node was introduced to a graph by identifying a transition using the transition estimation system. Entire utterances between the last transition point and its previous one were allocated to this node. A document

vector was produced for the node by applying the LDA vectorization scheme. It was then compared against all the previous nodes by computing their cosine similarity. Finally, the correspondence matrix  $C_{n \times n}$  was expanded gradually by concatenating the similarity scores of the new node. A separate experiment was conducted using the oracle *pre-segmented* transcriptions in order to investigate the effect of errors posed by the segmentation (new node detection) on the overall system performance.

The accuracy of an estimated correspondence matrix  $C_{n \times n}$  explains the quality of the final topological map. To compare an estimation  $C$  against its reference  $GT$ , ROC curves (Fawcett, 2006) were plotted for the estimation at each SNR. Since the matrix  $C$  is symmetric and its main diagonal does not reflect the estimation performance only the upper triangle (or lower) components of  $C$  (excluding its main diagonal) are required to be compared. A full range of thresholds (i.e. 0.00, 0.01, 0.02, ..., 1.00) was applied to the estimated values to be used for computing *True Positive Ratio* (TPR) and *False Positive Ratio* (FPR) for the ROC curve as:

$$TPR = \frac{TP}{TP + FN} \quad (4.3)$$

$$FPR = \frac{FP}{FP + TN} \quad (4.4)$$

where  $TP$ ,  $TN$ ,  $FP$ , and  $FN$  denote true-positive, true-negative, false-positive and false-negative respectively.

Figure 4.13 shows the ROC curves at different SNRs for each map-setting. ROC curves close to the dashed line of the diagonal indicate random estimations. For instance, the ROC curves presented in Fig-

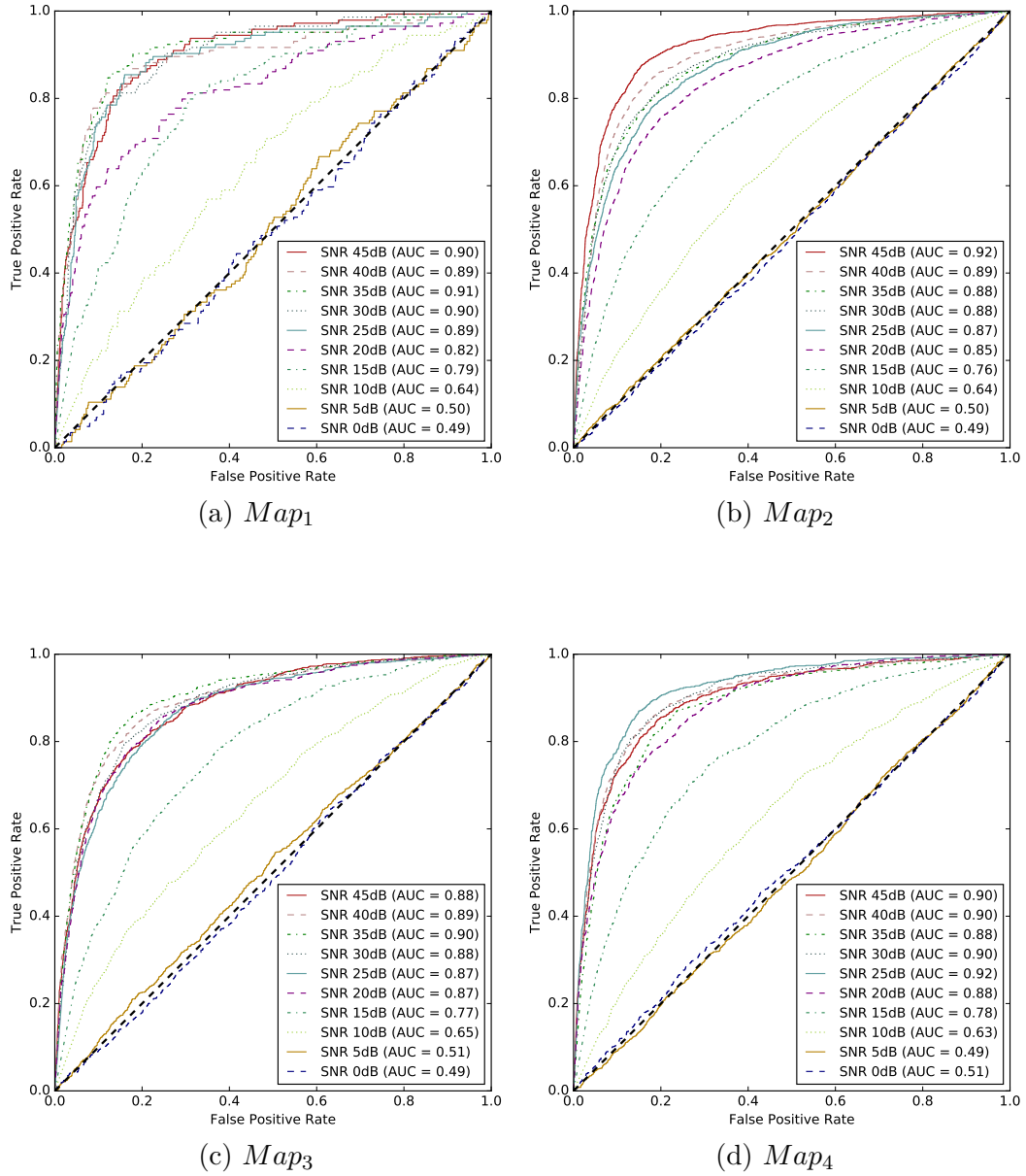


Figure 4.13 ROC curves at different SNRs for each map-setting. ROC curves close to the dashed line in the diagonal explain a random estimation.

ure 4.13 show that the estimated correspondence values at SNRs of 5 dB and 0 dB are randomly generated.

The AUC metric was used to score the performance of the system. The AUC of each map-setting is presented in Figure 4.14 as a function of the SNR. A perfect estimation receives an AUC of *one* and a random estimation is scored with a value close to 0.5. Figure 4.15 illustrates the overall AUC performance of the system as a function of WER. Results show there is no statistically significant difference in the system performance on transcriptions with about 16% and 25% WERs. The results also show that, while the performance reduces on WERs of greater than about 25% however, the system does not fail completely for error rates of about below 70%.

In comparison with the auto-segmented results, the pre-segmented system performance is plotted with a black line to illustrate the negative effect of node detection errors on the overall system performance. This explains that the node detection errors results in average 0.05 AUC performance reduction on WERs less than about 70%. For WERs less than 20%, these performance differences are barely significant ( $p=0.073$ ). However, the differences are statistically significant on WERs of about 20% to 70%.

It is notable that, while at extreme high WERs, the system performs randomly, the variance of the results is small. This low variance is explained by looking at the ASR transcripts. Under very low SNR conditions, and when the estimated acoustic likelihoods are not strong, ASR generates word sequences based on its language model. Such transcripts are not random. They are often full of high probability general function words with almost no sign of any important key-words. Thus, any segment of such text (as a node) looks highly similar to every other one. This can explain the reason why there is almost no difference between the pre-segmented and the auto-segmented results and also, the very low variance across different samples at such high WERs.

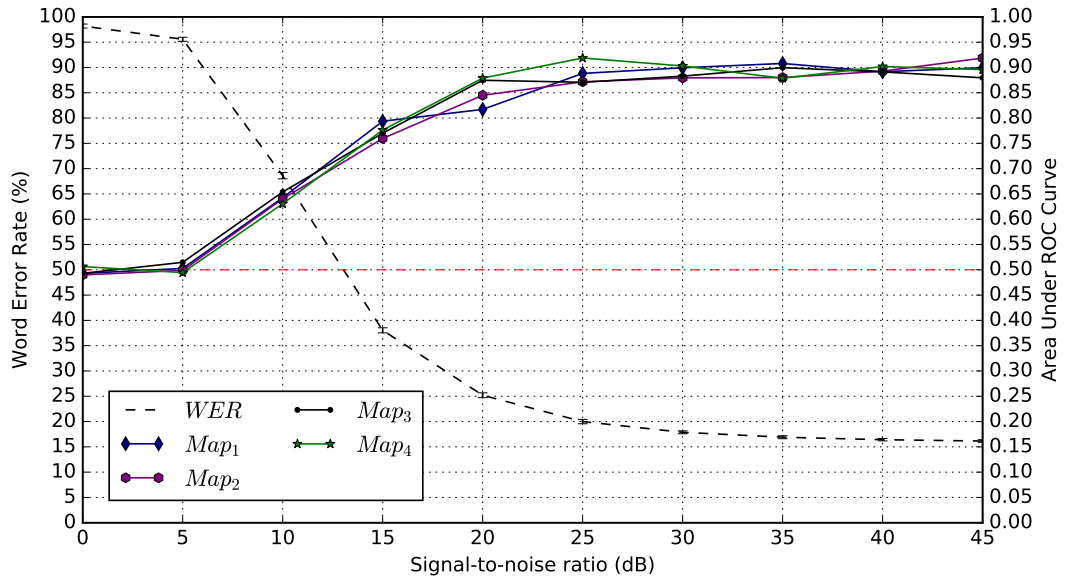


Figure 4.14 The ASR transcription WER on different SNRs is shown with dashed line. The AUC for each map-setting is presented as a function of SNR. Random estimation scores a value close to 0.5.

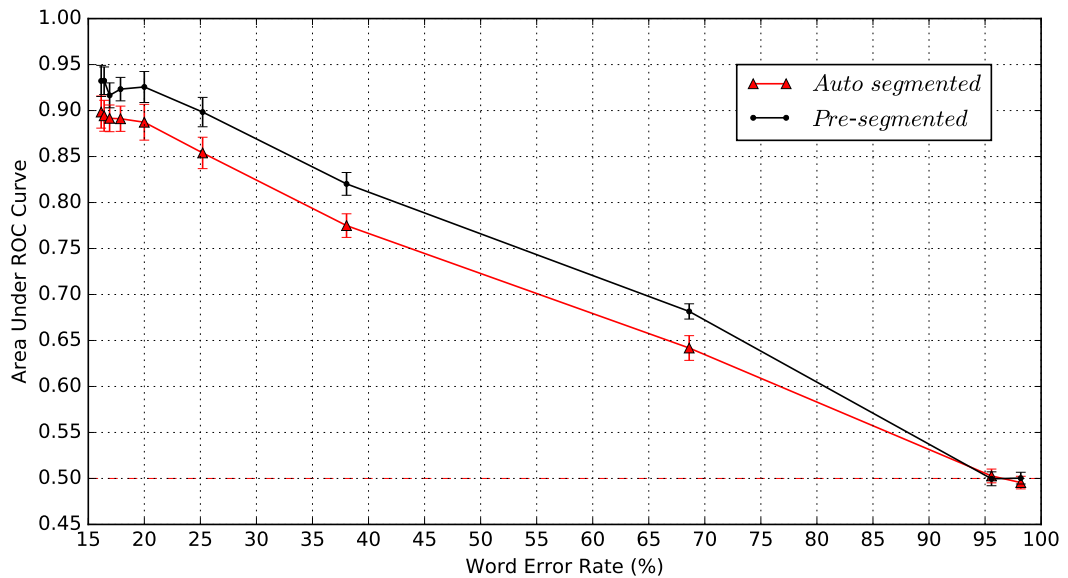


Figure 4.15 The overall AUC performance of the systems as a function of WER. The red line illustrates performance of the system with automatic segmentation. The black line represents the pre-segmented system performance.



## Summary

*Two novel approaches were presented for extracting locational information from a search and rescue speech conversation. Initially, a speech-based localization approach was introduced to estimate the location of a first responder in a finite set of location types. A topic-based perspective was employed since it offers the potential of being robust to high error rates in the automatic recognition of noisy speech. LDA vectorization technique was used to compare the semantic information of a first responder's spoken reports in a low-dimension feature space. A speech-based topological map estimation technique was then introduced that is inspired by automatic topological mapping algorithms. The new node insertion and correspondence estimation problems were framed as topic segmentation and text document similarity estimation tasks respectively. The impact of transcription errors on systems performance was investigated by experimenting on automatic transcripts of the SSAR corpus speech data with different SNRs. Results for both systems showed no significant performance decrease at a WER of about 25% compared with the performance on clean speech data with about 16.1% WER. Experiments results on transcriptions with WERs from about 25% to 70% demonstrated moderate performance declines compared with clean speech transcripts. The results demonstrated the feasible of using topic segmentation and identification techniques as foundations for developing systems to extract high-level information from natural conversations.*

# Chapter 5

## Design of a Two-Pass Speech Recognizer

---

The need for automatic recognition of conversational speech is emerging in a variety of application domains such as search and rescue (see Chapter 1). The automatic transcription of voice communication channels is difficult due to the high acoustic variations posed by the environment noise conditions, spontaneous speaking style, channel characteristics and the speakers' condition. In addition to these challenges, the statistical properties of the language often changes during a conversation due to change in context (see Section 2.2).

Under these circumstances, and especially when the estimated acoustic likelihoods are not so strong, an adaptive context-specific language model can play an important role in guiding the search for determining the most likely sequence of spoken words. High-level contextual information is often used to adapt a general language model depending on the ongoing situation (see Section 2.2.1.2). State-of-the-art ASR systems frequently obtain this information from a variety of additional sources, for instance mobile phone geolocation signals (Chelba et al., 2015), personalized user information (Wen et al., 2013) and domain information (Wen et al., 2013). Natural conversa-

tions often convey high-level information about the dialogue subject and context. It can be hypothesised that, such a high-level contextual information can be used along with other sources to improve the recognition of natural conversations.

Chapter 4 has shown that a topic-based perspective can be used to extract locational information from a highly imperfect automatic transcription of spoken conversations in a crisis response. This chapter presents a new two-pass speech recognition architecture which dynamically adapts its language model based on high-level location information content of a conversation.

## 5.1 System architecture

This section presents a two-pass speech decoding architecture with the objective of improving the transcription accuracy of the speech communication channels. Figure 5.1 visualises the main components of this system. The first decoding pass is an ASR search to determine the most likely sequences of spoken words given the decoder models. The language model used in the first decoding pass is a general static model trained on in-domain data. The first decoding pass produces its multiple word sequence hypotheses in a word lattice form. The generated lattices are retained to be rescored in the second pass. In addition, the best word lattice path (1-best hypothesis) is computed to be used by the speech-based *location identification* (location-ID) module.

Recalling from Section 4.1, the location-ID provides location estimations from highly inaccurate output of the ASR system. Relying on a topic-based approach and LDA vectorization principle, it estimates the speaker's location in some predefined room-types by tracking the



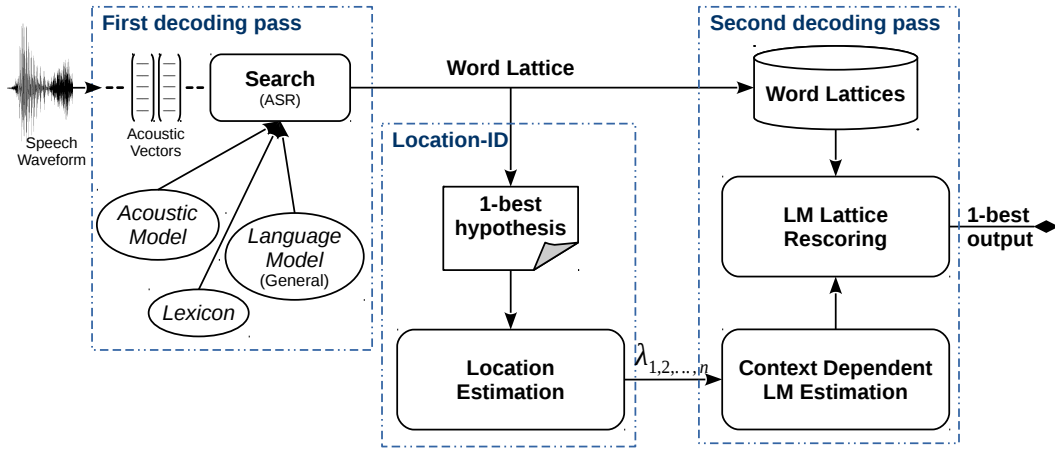


Figure 5.1 A general structure of the proposed two-pass speech decoding architecture. The location-ID module provides location estimations from highly inaccurate output of the ASR system. The second pass decoding stage is initiated whenever a new location is identified. All of the stored word lattices related to the recently identified location are then rescored based on the dynamically generated language model. The best path is computed for each word lattice as the final decoding hypothesis.

changes in the content of a report over time. The transition detection is constantly monitoring the 1-best decoding results to identify an estimated transition time. By detecting a transition point, the SVM class membership probability distribution for location classes is estimated for the recently identified segment to be used in the second pass decoding.

The second pass decoding stage is initiated whenever a new location is identified (i.e. after each transition detection). A context-dependent language model is built for each identified location by combining a collection of location-specific language models according to their relevance obtained from the location-ID. The models were combined via the mixture interpolation (see Section 2.2.1.3). The interpolation coefficients ( $\lambda_i \ \forall i \in [1, 2, \dots, n]$  for  $n$  location-specific

models) were set based on the SVM location class membership probability distribution of the location-ID. Based on the language model lattice rescoring described in Section 2.2.3.1, the dynamically generated model is then used to rescore all the stored word lattices related to the latest identified location. Finally, the best path is computed for each word lattice for the final decoding hypothesis.

## 5.2 Experimental set-up

The SSAR corpus (described in Chapter 3) was used to evaluate the proposed system on the task of transcribing voice communication channels. The entire dataset was divided based on a K-Fold cross-validation scheme ( $K=12$ ) into training, test and development datasets with about 10h, 1h, and 1h of data respectively (similar to the previous chapter experiments).

### 5.2.1 Baseline automatic speech recognition system

The same speech decoder described in the previous chapter was used as the baseline recognition system. The particular system acoustic models were trained on about 10 hours of clean speech data in the training subset using the Kaldi open-source speech recognition toolkit (Povey et al., 2011). These models were GMM/HMM systems with 13-dimensional MFCCs with  $\Delta$  and  $\Delta\Delta$  (deltas and accelerations respectively). MFCCs were spliced across three frames of left and right context and reduced to a 40 feature vector using linear discriminant analysis whose class is one of 2500 tied triphone HMM states. The tied states are modelled by a total of 15000 Gaussians. On top of that, the specific system was trained on adapted features using *Maximum Likelihood Linear Transformation* (MLLT) (Gopinath, 1998),

and *Feature-space Maximum Likelihood Linear Regression* (fMLLR) (Li et al., 2002) with *Speaker Adaptive Training* (SAT) (Anastasakos et al., 1996).

For decoding a 16K lexicon, a 3-gram language model was built by the linear interpolation of two models (see Section 2.2.1.3). The background language model is a relatively large model trained on the transcriptions of the Switchboard telephone speech corpus (Godfrey et al., 1992) with about 3 million words. The second model was trained on the training subset annotations with about 65k words. Both of the language models were trained and interpolated using the SRILM toolkit (Stolcke, 2002). The interpolation weight of these two ( $\lambda_b$ ) was tuned by finding the minimum perplexity of the interpolated language model on the held-out validation dataset. The best interpolation coefficient of  $\lambda_b=0.4$  was determined and applied throughout all subsequent experiments. The pronunciations lexicon used in constructing the speech recognition system was the BEEP dictionary (Robinson, 1996).

### 5.2.2 Experiments

Three sets of experiments were conducted on the speech data with different levels of artificially added noise to the test and development sets:

1. The first set of experiments was for providing a baseline recognition performance using a typical ASR system. The baseline ASR system was used to decode the test-sets and the transcription WER was computed.
2. The second set of experiments investigated the effect of employing location information for language model lattice rescoring in

the described system. These experiments utilized the actual location transcriptional tags (*oracle-condition*).

3. Finally in the third set of experiments, the *Estimated Location Information* (ELI-condition) was obtained from the initial recognition results. This estimated result was used for dynamically generating the lattice rescoring language model.

Similar to the experiments in the previous chapter, test and development datasets with noise levels ranging from clean speech to an SNR of 0 dB were built by mixing the clean utterances into the background noise (the CHiME-3 café noise). The experiments for each SNR followed a 12-fold cross-validation scheme. For each fold, the ASR acoustic models were trained on its clean train-set, and the best *Language Model Scaling Factor* (LMSF) and word insertion penalty (Jurafsky and Martin, 2009, chapter 9) were estimated on the development-set for each SNR. The trained ASR system was used for decoding all pre-segmented conversation utterances of the test-sets. The generated lattices for each utterance was stored to be processed after a location were identified. The best lattice path hypothesis was computed to be used by the location-ID.

The location-ID SVM classifiers were trained on the training dataset (See Section 4.1). By identifying a new transition point, the SVM class membership probability distribution for 13 *RoomType* classes was estimated for the recently identified segment to be used as language model interpolation coefficients ( $\lambda_{1-13}$ ) in building the context-dependent model. Figure 5.2 illustrates the process of building a context-dependent language model.

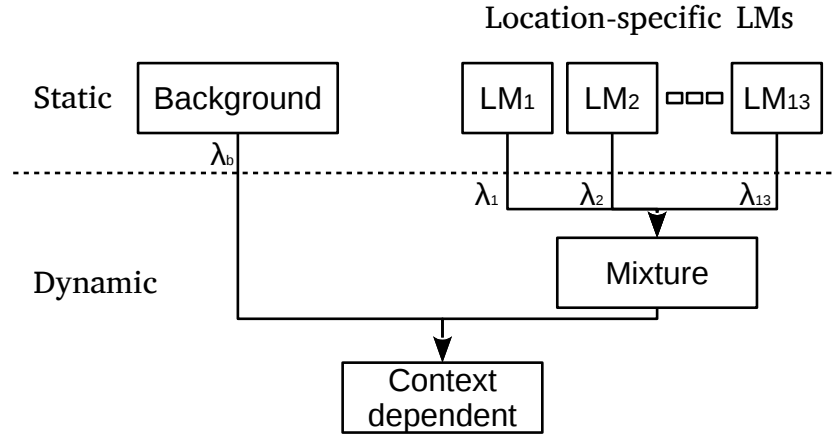


Figure 5.2 The process of building the context-dependent language model in the rescoring pass. The background language model is a large model trained on out-of-domain data (Switchboard corpus transcriptions) and the location-specific language models were small but specific models trained on location-specific collections of utterances in the training data. A context-dependent model was built by dynamically interpolating these static models based on the estimated  $\lambda_{1-13}$  coefficients.

The background language model used was the same large model trained on the transcriptions of the Switchboard telephone speech corpus. Each location-specific language model is a small but specific model trained on a collection of utterances in the training data which were related to a particular *RoomType*. The estimated interpolation coefficients  $\lambda_{1-13}$  by the location-ID were used to dynamically build a mixture model for each identified segment of a conversation. This mixture model was then interpolated with the background model making the final context dependent language model.

The context dependent language model was then plugged into the rescoring of a word lattice. The rescoring was performed in the Kaldi ASR toolkit by initially decoupling the scores of acoustic and language models of a lattice's transition arcs and then replacing the previous language model probabilities with the new language model

probabilities. The rescored lattice was used to find a new one-best transcription hypothesis as the output of the described system.

### 5.3 Results and discussion

This section presents the results of the experiments and compares the systems performance based on their WER. Figure 5.3 illustrates the baseline system WERs on different noise levels. More detailed information about the performance of the baseline system, including its word insertion, deletion, and substitution ratio on different SNRs, is also presented. Results show that the baseline system is capable of decoding the clean version of the speech signal with about 16.1% WER. In all SNRs the word insertion rate was small (about 3%) in comparison with the substitution and deletion rate. This can

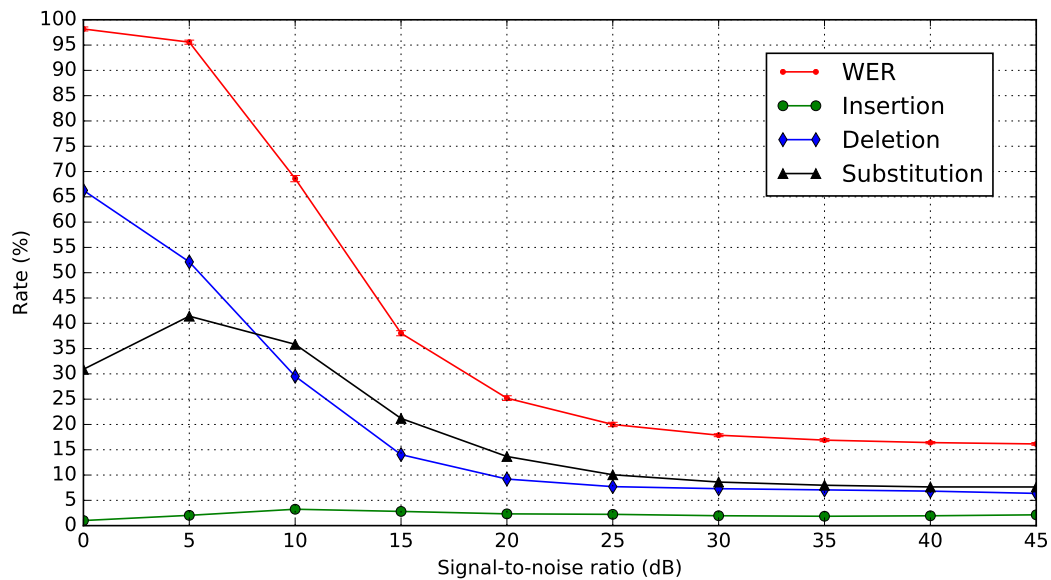


Figure 5.3 The detailed information about the performance of the baseline system, including its word insertion, deletion, and substitution ratio on different SNRs.

be related to the speaking rate of conversational speech since high speaking rate has been reported to result in low speech insertion and high deletion and substitution errors (Martinez et al., 1998; Nanjo and Kawahara, 2004). The low insertion rate can also be related to the word insertion penalty that was tuned on the development-set for each SNR. The particular effect of these parameters on the word insertion rate is not investigated within the scope of this thesis. In SNRs of about 20 dB, 15 dB and 10 dB, the substitution rate was about 25% higher than the deletion rate. This shows that the ASR acoustic model mismatches with the noisy speech signal and wrong words were hypothesised based on the language model. At extreme SNRs of about 5 dB and 0 dB, a majority of the words were deleted because the ASR failed to detect any speech in such low SNRs.

To compare the performance of the baseline ASR with the system in the ELI-condition, Figure 5.4 shows their WER on different SNRs. The performance of the location-ID module is also plotted in this

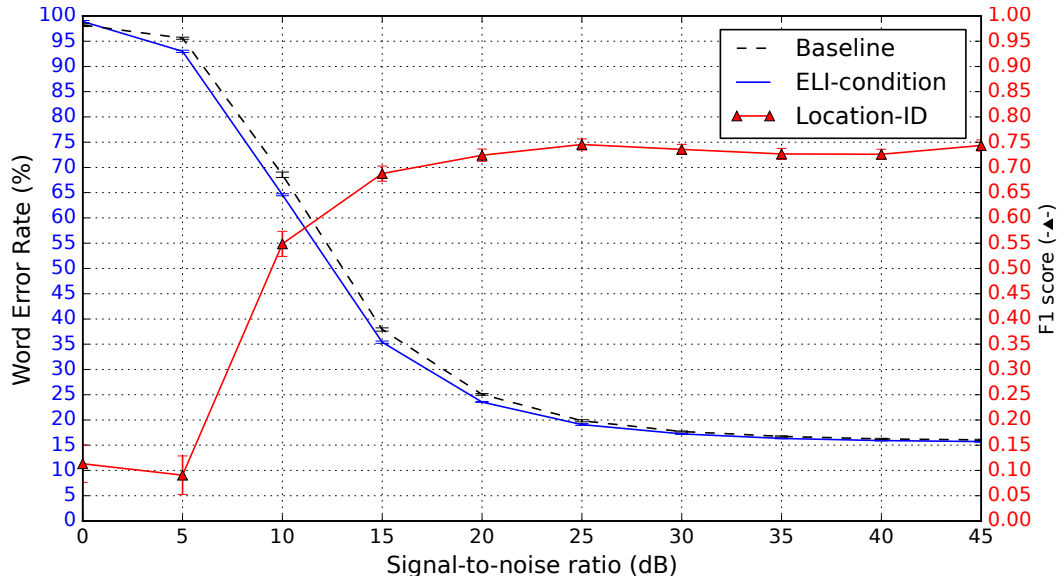


Figure 5.4 WER of the baseline ASR system compared with the ELI-condition system as a function of SNR increase. Performance of the location-ID module as a function of SNR is also presented.

figure. The recognition errors rise as a result of decreasing SNR in both baseline and ELI-condition. It is notable that the system in ELI-condition has performed better compared to the baseline system with moderately lower WER in all SNRs.

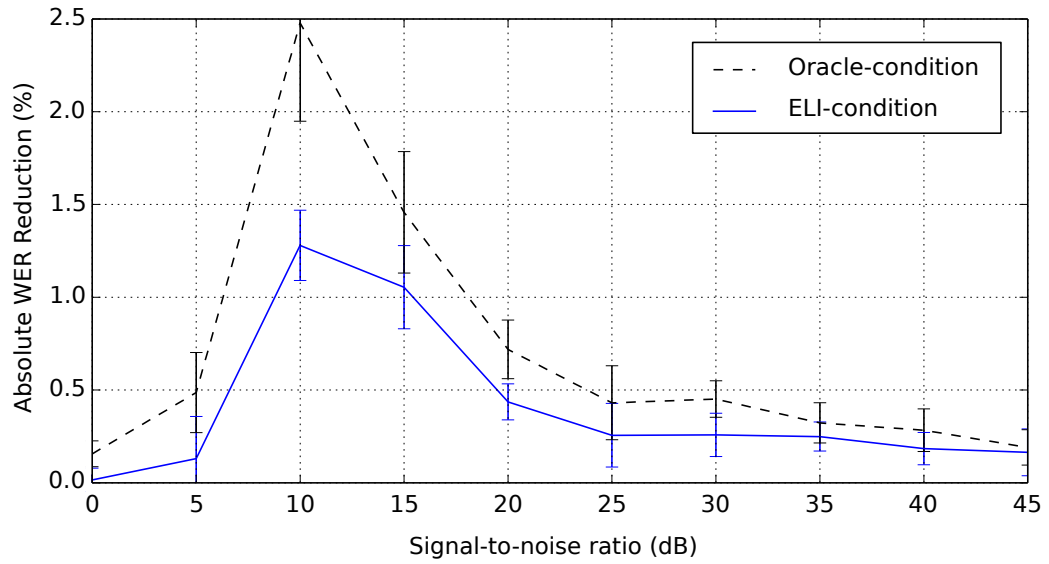
To illustrate this performance difference clearly, Figure 5.5a shows the absolute WER reduction values (transcription accuracy gain) for both the oracle and the ELI conditions. Figure 5.5b presents these WER reductions relative to the baseline ASR. These graphs reveal that these transcription accuracy gains start from a small, but statistically significant values from about 1% relative at clean speech and peaks at about 3% relative in an SNR of 15 dB. The statistical significance of these WER reductions are tested with the matched-pairs test (Gillick and Cox, 1989) with a  $p\text{-value} < 0.05$ .

To investigate the general gain increase in lower SNRs, Figure 5.6 illustrates the LMSF value as a function of the SNR. This graph shows a decline on the LMSF in low SNRs. The LMSF, which is an exponent on the language model probability,  $P(W)$ , has the effect of decreasing the value of language model probability (since the  $P(W)$  is less than one and the LMSF is greater than one) (Jurafsky and Martin, 2009, chapter 9):

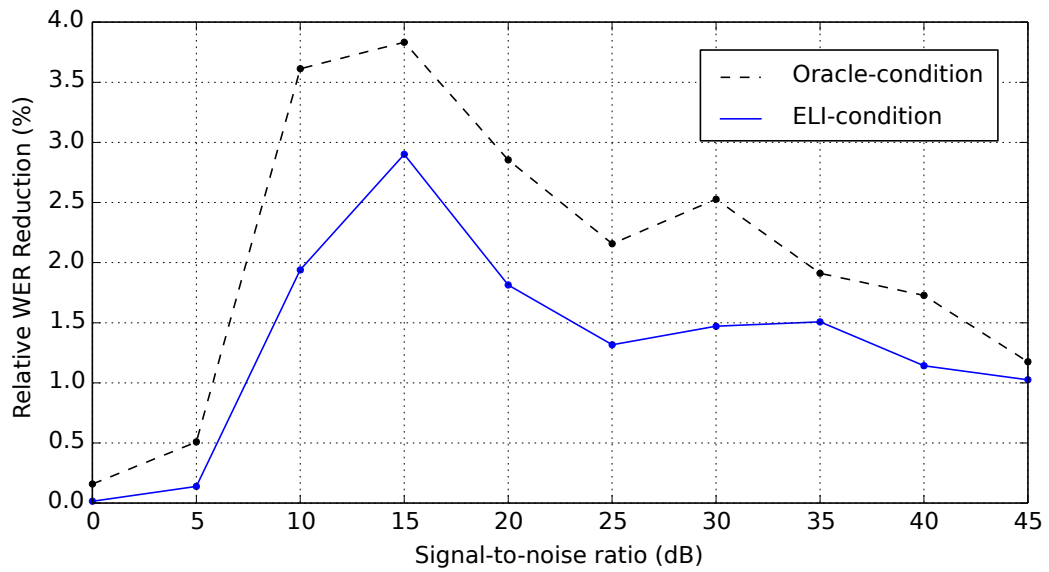
$$\hat{W} = \operatorname{argmax}_{W \in \mathcal{L}} P(O|W)P(W)^{LMSF} \quad (5.1)$$

The LMSF reduction reveals the decoder tendency of moving towards more weight on the language model in low SNRs (see Figure F.1 and Figure F.2 in Appendix G for the baseline ASR performance at different SNRs and all LMSFs). This explains the general gain increase in higher levels of noise when a context-specific language model has a greater effect and when the acoustic likelihoods are not so strong. This can also explain the modest or lack of recognition improvements





(a) Absolute WER reduction



(b) WER reductions relative to the baseline

Figure 5.5 (a) The absolute WER reductions for the system in both oracle and ELI-conditions. (b) The WER reductions relative to the baseline ASR for the system in both oracle and ELI conditions.

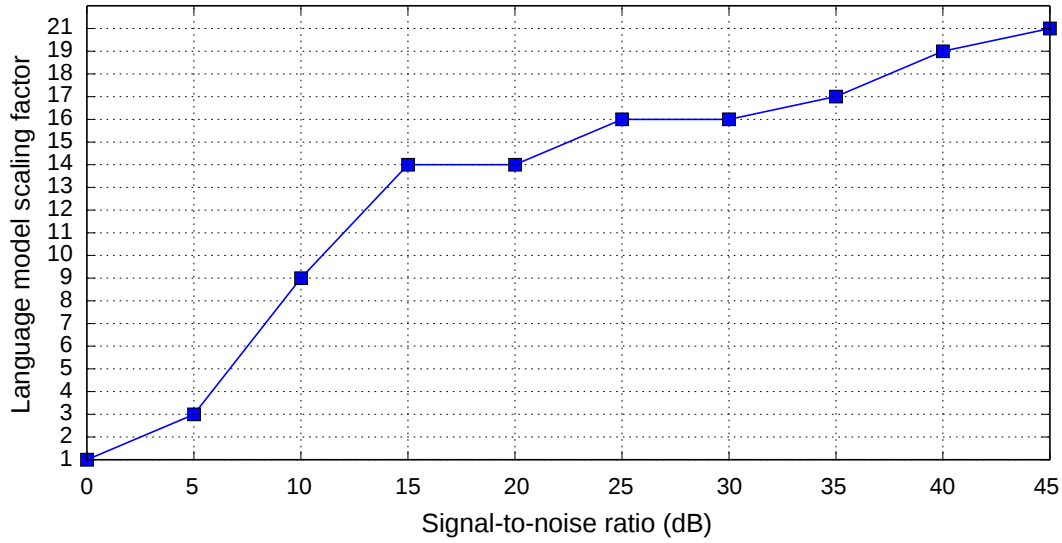


Figure 5.6 *Estimated LMSF on the development-sets for each SNR.*

in some previous studies such as Chelba et al. (2015) even after observing a large perplexity reduction by using a specific language model.

Figure 5.5b shows that in both the oracle and ELI condition set-ups, their gains decrease at SNRs lower than about 15 dB. Extremely low acoustic likelihoods at high noise levels explain these low gains. In other words, with the absence of a reliable observation and pure reliance on the language model prior, the ASR hypotheses are not close to their reference. Consequently, the language model adaptation cannot offer improvements in these conditions. In comparison with the oracle condition, the ELI gain is lower on SNRs of 15 dB and 10 dB. This has an additional reason which is the failure of the location-ID module in extreme WERs of the initial transcripts (see location-ID performance in Figure 5.4).



### Summary

*This chapter has presented a new two-pass speech recognition architecture for transcribing voice communication channels within a crisis response context. The speech-based location estimation system (cf. Section 4.1) was employed to extract high-level information about the location context from imperfect transcription output of the first decoding pass. The estimated information was then used to dynamically update interpolation coefficients for building a mixture model of  $n$  context-specific language models. The context-specific model was used for rescoring the initial word lattice in a second decoding stage. The experiments on the SSAR corpus reported a modest but statistically significant WER reduction on clean speech. Experiments on different noise levels showed relatively higher word error reductions in an SNR of about 15 dB compared with that on clean speech data. Experiments on SNRs of about 5 dB and 0 dB were resulted in a complete system failure because of the extremely low acoustic likelihoods as well as the location-ID module collapse in the extreme WERs.*



## Conclusions

---

This chapter summarises the reported research and draws a conclusion with respect to the research questions laid out in Chapter 1. It also provides an outlook on a number of potential directions for future work.

### 6.1 Reviewing the scope of the thesis

This research was motivated by the need for automated solutions for extracting valuable information from speech communications within application domains such as search and rescue. The primary aim of this thesis was to investigate the feasibility of developing a system for estimating first responder's location and the incident scene layout by discovering the context and content of voice communication channels.

The following steps were identified at the outset as necessary milestones to reach the target:

- Identify the major challenges and limitations in automatic processing of speech communication channels in a search and rescue domain.

- Survey the related background issues and the state-of-the-art in processing natural speech conversations.
- Provide an appropriate speech dataset comprising task-related annotated conversations by targeting the goals and needs of the information extraction task in the context of crisis response.
- Development and evaluation of topic-based locational information extraction in the context of simulated search and rescue communications.
- Investigate the utility of exploiting the extracted high-level locational information for improving speech recognition performance.

A brief overview of search and rescue voice communication systems was provided in Section 2.1. The role of speech technology in accessing their information content for situational awareness formation was then discussed. Focusing on the difficulty of an automatic speech recognition task, the voice and language parameters in this domain were characterized. The related backgrounds in speech recognition were presented briefly in Section 2.2. In Section 2.3, state-of-the-art in topic detection approaches were studied in more detail.

To provide a suitable speech dataset in the task of locational information extraction from speech communications, a new goal-oriented conversational speech corpus was designed and collected (cf. Chapter 3). The SSAR corpus was recorded based on an abstract communication model during a search process in a simulated crisis response training scenario. Each conversation is concerned with a cooperative task of exploring a simulated indoor environment by a first responder and estimating a topological map by a task leader. The SSAR corpus

comprises of 96 dialogues between 24 speakers, totalling 12 hours, with 80K words transcribed manually. Aligned with these recordings, other information about the participants' locations, actions and objects in their field of view in the environment are available in computer readable log-files. This information can be used as a form of conceptual annotation for the conversations. It is anticipated that 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 automatic speech recognition. This corpus is being made available for research purposes (via LDC).

Two novel approaches were presented for extracting locational information from a search and rescue speech conversation (cf. Chapter 4). Initially, a speech-based localization approach was introduced to estimate the location of a first responder in a finite set of location types. A topic-based perspective was employed since it offers the potential of being robust to high error rates in the automatic recognition of noisy speech. An LDA vectorization technique was used to compare the semantic information of spoken reports in a low-dimension feature space. A speech-based topological map estimation technique was then introduced that is inspired by automatic topological mapping algorithms. The new node insertion and correspondence estimation problems were framed respectively as topic segmentation and text documents similarity estimation tasks.

The proposed systems were evaluated on the SSAR corpus. The impact of transcription errors on systems performance was investigated by experimenting on automatic transcripts of the SSAR corpus speech data with different SNRs. Results for both systems demonstrated no significant performance decrease at a WER of about 25% compared with the performance on clean speech data with about 16.1% WER.

Experimental results on transcriptions with WERs from about 25% to 70% demonstrated moderate performance declines compared to that obtained on the transcription of clean speech data.

Of course presented results should be viewed with caution in considering their application to real search and rescue. This is because the proposed systems were tested on a conversation dataset which is collected based on a simulated task of exploring a limited number of indoor environments. In a real scenario, location types can be more diverse which may influence the accuracy of the location identification system. Furthermore, an incident environment may not be well structured and can change during the mission time. Therefore, the dynamics of an incident scene also needs to be taken into the account for map estimation. Nevertheless, the described systems and their performance demonstrated the feasibility of using topic segmentation and identification techniques as foundations for developing systems to extract high-level information from natural conversations.

The final milestone was achieved by investigating the utility of exploiting the extracted high-level location information for improving speech recognition performance. A new two-pass speech decoding architecture was presented (cf. Chapter 5). The location estimation from a first decoding pass was used to dynamically adapt a general language model and rescore the initial recognition hypotheses. The experiments on the SSAR corpus reported a modest but statistically significant WER reduction on clean speech. Experiments on different noise levels showed the highest WER reduction of about 3% (relative to the baseline) on an SNR of about 15 dB. A similar experiment with oracle location information showed the highest WER reduction of about 4% (relative to the baseline) on the same SNR of about 15 dB.



The location identification module is a fundamental component of the described system. Any improvement in the location estimations can potentially result in higher recognition gains up to the oracle-condition performance. One drawback of this architecture is that the location identification module computes a new estimation after a speaker has moved to a new location (when a new segment is identified). This means the location identifications, and subsequently, the speech recognitions are not in real-time.

## 6.2 Answer to research questions

The first research question was: can topic detection techniques be used to derive high-level information (such as location information) from speech communication channels in a search and rescue environment? It has been shown (in Chapter 4) that topic detection techniques can be successfully used to extract high-level locational information from speech communication channels in a task of exploring and describing a simulated environment. Looking from a topic-based perspective and tracking the changes in the content of a spoken report provided an estimation of first responders' location. The new node insertion and correspondence estimation problems were successfully framed as a topic segmentation and tracking task.

The second research question was: can such high-level information be used top-down to improve speech recognition performance? It has been shown (in Chapter 5) that a statistically significant improvement can be obtained on noisy speech by exploiting the estimated location information for rescored ASR recognition hypotheses.

### 6.3 Original contributions

The main scientific contributions resulting from the research reported in this thesis are as follows:

- Successfully demonstrated the feasibility of using topic detection techniques for extracting high-level locational information in a simulated search and rescue context.
- Designed and evaluated a novel topic-based approach for location estimation in a simulated search and rescue context.
- Designed and evaluated a novel topic-based approach for topological map estimation in a simulated search and rescue context.
- Provided experimental evidence that a statistically significant improvement can be obtained on noisy speech by exploiting the estimated location information for rescored an ASR recognition hypotheses.
- Designed and evaluated a new two-pass speech decoding architecture for transcribing the voice communication channels within the search and rescue context.
- Developed a new goal-oriented conversational speech corpus for tasks of human/human conversation understanding and automatic speech recognition.

### 6.4 Future work

Evidence gathered throughout this research has shown the feasibility of using a topic-based perspective for developing systems to extract

high-level situational information from conversations in a search and rescue domain. Nevertheless, the findings lead to new questions and requirements for further improvements which are considered as future research directions as follows:

- The presented speech-based location estimation system introduced a new source of information to the field of localization. One outstanding issue, which should be investigated in future studies, is the potential for integration of this system with other localization techniques such as *Simultaneous Localization and Mapping* (SLAM) (Burgard and Hebert, 2008). This integration can provide a strong multimodal approach for the location estimation task.
- Another issue which is not investigated in this research is the dynamic nature of a search and rescue environment. Future investigations are required to model the dynamics of an incident scene and first responders knowledge of an ongoing situation.
- Future studies can investigate the utility of exploiting the rich information content of ASR lattice outputs. Instead of the single-best ASR hypothesis, using lattices along with their information about word confidence scores has been reported to obtain performance increase in SLU tasks such as named entity extraction and call classification (Hakkani-Tür et al., 2006). Adoption of similar strategies can add to the robustness of the presented systems.
- Acoustic information is what distinguishes speech processing from text analytics. Acoustic and prosodic features can be used

in combination with features extracted from ASR outputs for a more accurate locational information extraction.

- The presented topological map estimation system relies on the information content of a single speech conversation. However, in real scenarios, a number of first responders explore an incident scene. There are potential opportunities for processing multiple parallel conversations. Integration of all communication channels can provide a complete view of the incident scene layout. Overlaps between multiple observations can improve the map estimations. Prior knowledge about the environment (e.g. an architectural map) may also improve the identification performance by limiting the search space for the location of a first responder.
- A more sophisticated speech decoding system can benefit from the integration of different location information sources. Figure 6.1 illustrates an envisaged system architecture. In this architecture, a speech-based localization and mapping module can provide estimations from a combination of speech and other information sources (e.g. an architectural map of a building or a map that has been generated by rescue robots). Extraction of locational information from speech can contribute to updating the locational information that has been gathered during the time of a search and rescue mission. The collected information about an environment map and the latest information about first responders' location can provide estimates of individuals' location at anytime. A second decoding pass can employ this information and estimations to dynamically generate a location-specific language model to be used in rescoreing the generated

word lattices. This may address the mentioned real-time issue of the introduced speech decoding system.

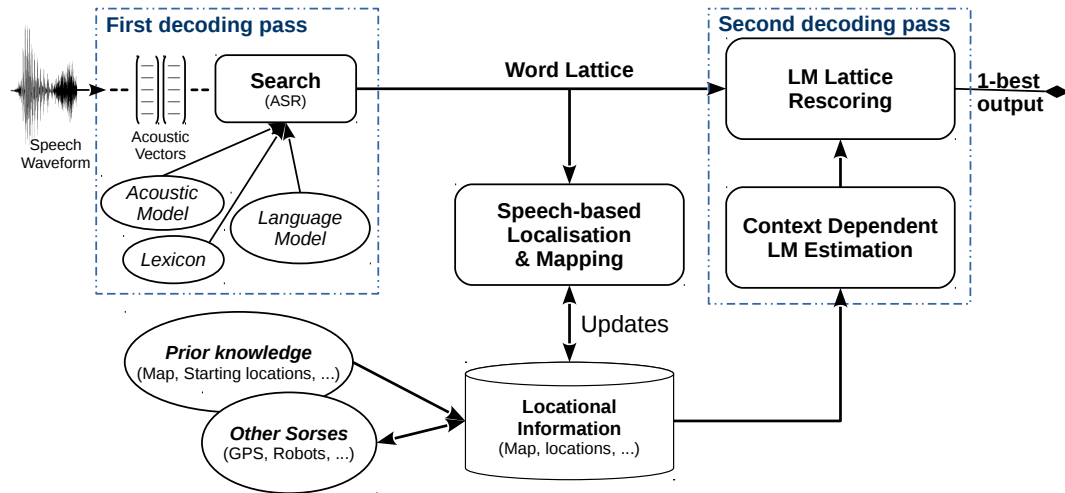


Figure 6.1 *An envisaged speech decoding architecture. The speech-based localization and mapping module can provide its estimations from a combination of speech and other information sources. These estimations can be used to contribute in updating the locational information gathered during the time of a search and rescue mission. The second decoding pass can use the collected information to dynamically generated a location-specific language model to be used in rescoring the generated word lattices. The best path of each word lattice is the final decoding hypothesis.*

## 6.5 Conclusion

Access to the information content of speech conversations is important for situation awareness formation in any search and rescue operation. Automatic systems can be introduced to help the current support systems by extracting critical information from all conversations. The achievements presented in this thesis demonstrated the feasibility of using topic segmentation and identification techniques as foundations for developing systems to extract high-level informa-

tion from natural conversations. The proposed systems for locational information extraction exhibit to be robust to an imperfect transcription of spontaneous and noisy speech. It has also been shown that the derived information can be used for improving speech recognition performance. These findings can be useful to future investigations on processing communication channels in search and rescue missions. They can also be beneficial for formulating the design of practical systems in this domain as well as similar application areas.

## References

---

- Allan, J., Carbonell, J. G., Doddington, G., Yamron, J. and Yang, Y. (1998), Topic detection and tracking pilot study final report, *in* ‘Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA)’.
- Allan, J., ed. (2012), *Topic detection and tracking: event-based information organization*, Vol. 12, Springer Science & Business Media, New York, NY, USA.
- Allen, J. and Heeman, P. (1995), ‘TRAINS Spoken Dialog Corpus LDC95S25’, CD.
- Anastasakos, T., McDonough, J., Schwartz, R. and Makhoul, J. (1996), A compact model for speaker-adaptive training, *in* ‘Proceedings of International Conference on Spoken Language Processing (ICSLP)’, Vol. 2, pp. 1137–1140.
- Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J. and Miller, J. (1991), ‘The HCRC map task corpus’, *Language and Speech* **34**(4), 351–366.
- Androutsopoulos, I., Koutsias, J., Chandrinou, K. V. and Spyropoulos, C. D. (2000), An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages, *in* ‘Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval’, ACM, pp. 160–167.
- Ang, J., Liu, Y. and Shriberg, E. (2005), Automatic Dialog Act Segmentation and Classification in Multiparty Meetings, *in* ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 1061–1064.

- Angeli, A., Filliat, D., Doncieux, S. and Meyer, J. A. (2008), ‘Fast and incremental method for loop-closure detection using bags of visual words’, *IEEE Transactions on Robotics* **24**(5), 1027–1037.
- Arguello, J. and Rosé, C. (2006), Topic segmentation of dialogue, *in* ‘Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech’, Association for Computational Linguistics, pp. 42–49.
- Ashish, N., Eguchi, R., Hegde, R., Huyck, C., Kalashnikov, D., Mehrotra, S., Smyth, P. and Venkatasubramanian, N. (2008), Situational Awareness Technologies for Disaster Response, *in* H. Chen, E. Reid, J. Sinai, A. Silke and B. Ganoz, eds, ‘Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security’, Springer, pp. 517–544.
- Aubert, X. and Ney, H. (1995), Large vocabulary continuous speech recognition using word graphs, *in* ‘IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, Vol. 1, pp. 49–52 vol.1.
- Audacity (2016), ‘Audacity: A free, open source software for recording and editing sounds’.
- URL:** <http://www.audacityteam.org/>
- Banerjee, S., Rose, C. and Rudnický, A. I. (2005), The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing, *in* ‘IFIP Conference on Human-Computer Interaction’, Springer, Berlin, Heidelberg, pp. 643–656.
- Barker, J., Marxer, R., Vincent, E. and Watanabe, S. (2016), The CHiME challenges: Robust speech recognition in everyday environments, *in* ‘New era for robust speech recognition - Exploiting deep learning’, Springer.
- Barnett, J., Anderson, S. W., Broglio, J., Singh, M., Hudson, R. and Kuo, S. W. (1997), Experiments in spoken queries for document retrieval, *in* ‘Proceedings of Eurospeech’, Citeseer, ISCA.
- Beeferman, D., Berger, A. and Lafferty, J. (1999), ‘Statistical models for text segmentation’, *Machine learning* **210**(1-3), 177–210.
- Bellegarda, J. R. (2004), ‘Statistical language model adaptation: Review and perspectives’, *Speech Communication* **42**(1), 93–108.



- Betts, B. J., Binsted, K. and Jorgensen, C. (2006), ‘Small-vocabulary speech recognition using surface electromyography’, *Interacting with Computers* **18**(6), 1242–1259.
- Bird, S., Klein, E. and Loper, E. (2009), *Natural language processing with Python*, O’Reilly Media, Inc.
- Blei, D. M. and Moreno, P. J. (2001), Topic segmentation with an aspect hidden Markov model, *in* ‘Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval’, ACM, pp. 343–348.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), ‘Latent Dirichlet Allocation’, *Journal of Machine Learning Research* **3**, 993–1022.
- Boal, J., Sánchez-Miralles, Á. and Arranz, Á. (2014), ‘Topological simultaneous localization and mapping: a survey’, *Robotica, Cambridge University Press* **32**(5), 803–821.
- Brooks, R. A. (1990), ‘Elephants don’t play chess’, *Robotics and Autonomous Systems* **6**(1), 3–15.
- Burgard, W. and Hebert, M. (2008), Simultaneous Localization and Mapping, *in* B. Siciliano and O. Khatib, eds, ‘Springer Handbook of Robotics’, Springer, chapter E-37, pp. 853–870.
- Canavan, A., Graff, D. and Zipperlen, G. (1997), ‘CALLHOME American English Speech LDC97S42’, DVD.
- Canavan, A. and Zipperlen, G. (1996), ‘CALLFRIEND American English-Non-Southern Dialect LDC96S46’, Web Download.
- Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D. and Wellner, P. (2006), The AMI Meeting Corpus: A Pre-announcement, *in* ‘International Workshop on Machine Learning for Multimodal Interaction’, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 28–39.

- Chelba, C., Zhang, X. and Hall, K. (2015), Geo-Location for Voice Search Language Modeling, *in* ‘Proceedings of Interspeech’, ISCA.
- Chen, L., Gauvain, J. L., Lamel, L. and Adda, G. (2003), Unsupervised language model adaptation for broadcast news, *in* ‘Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, Vol. 1, pp. 1–220.
- Chen, L., Gauvain, J.-L., Lamel, L., Adda, G. and Adda-Decker, M. (2001), Using information retrieval methods for language model adaptation, *in* ‘Proceedings of Interspeech’, ISCA, pp. 255–258.
- Chen, S. F. and Goodman, J. (1996), An empirical study of smoothing techniques for language modeling, *in* ‘Proceedings of the 34th annual meeting on Association for Computational Linguistics’, Association for Computational Linguistics, pp. 310–318.
- Chen, X., Tan, T., Liu, X., Lanchantin, P., Wan, M., Gales, M. J. F. and Woodland, P. C. (2015), Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition, *in* ‘Proceedings of Interspeech’, ISCA.
- Chien, J. T. and Chueh, C. H. (2011), ‘Dirichlet Class Language Models for Speech Recognition’, *IEEE Transactions on Audio, Speech, and Language Processing* **19**(3), 482–495.
- Choi, F. Y. Y. (2000), Advances in domain independent linear text segmentation, *in* ‘Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference’, Association for Computational Linguistics, USA, p. 8.
- Cieri, C., Miller, D. and Walker, K. (2004), 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, pp. 69–71.
- Claveau, V. and Lefèvre, S. (2015), ‘Topic segmentation of TV-streams by watershed transform and vectorization’, *Computer Speech & Language* **29**(1), 63–80.

- CMU (1998), *The Carnegie Mellon University (CMU) American English pronunciation dictionary, release 0.6*, Carnegie Mellon University.  
**URL:** <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>
- Dauids, A. (2002), ‘Urban search and rescue robots: From tragedy to technology’, *IEEE Intelligent Systems and Their Applications* **17**(2), 81–83.
- Davis, S. and Mermelstein, P. (1980), ‘Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences’, *IEEE transactions on acoustics, speech, and signal processing* **28**(4), 357–366.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990), ‘Indexing by latent semantic analysis’, *Journal of the American society for information science* **41**(6), 391.
- Dehmer, M. and Emmert-Streib, F. (2008), ‘Structural information content of networks: Graph entropy based on local vertex functionals’, *Computational Biology and Chemistry* **32**(2), 131–138.
- Doddington, G. (1998), The Topic Detection and Tracking Phase 2 (TDT-2) Evaluation Plan: Overview & Perspective, *in* ‘Proceedings of the Broadcast News Transcription and Understanding Workshop (Sponsored by DARPA)’, Lansdowne, Virginia, pp. 223–229.
- Doulaty, M., Saz, O., Ng, R. W. M. and Hain, T. (2016), Automatic Genre and Show Identification of Broadcast Media, *in* ‘Proceedings of Interspeech’, ISCA.
- Echeverry-Correa, J. D., Ferreiros-López, J., Coucheiro-Limeres, A., Córdoba, R. and Montero, J. M. (2015), ‘Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition’, *Expert Systems with Applications* **42**(1), 101–112.
- Endsley, M. R. (1995), ‘Toward a Theory of Situation Awareness in Dynamic Systems’, *The Journal of the Human Factors and Ergonomics Society* **37**(1), 32–64.
- Evermann, G., Chan, H. Y., Gales, M. J. F., Jia, B., Mrva, D., Woodland, P. C. and Yu, K. (2005), Training LVCSR systems on thousands of hours

- of data, in ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 209–212.
- Fawcett, T. (2006), ‘An introduction to ROC analysis’, *Pattern Recognition Letters* **27**(8), 861–874.
- Fiscus, J. G. and Doddington, G. R. (2002), Topic detection and tracking evaluation overview, in J. Allan, ed., ‘Topic detection and tracking’, Springer, chapter 2, pp. 17–31.
- FLAME-SIM (2016), ‘FLAME-SIM: Fire department training simulation software’.
- URL:** <http://www.flame-sim.com>
- Frederic Bechet (2011), Named Entity Recognition, in G. Tür and R. De Mori, eds, ‘Spoken Language Understanding: Systems for Extracting Semantic Information from Speech’, Wiley, chapter 10, pp. 257–290.
- Fügen, C., Wölfel, M., McDonough, J. W., Ikbāl, S., Kraft, F., Laskowski, K., Ostendorf, M., Stüker, S. and Kumatani, K. (2006), Advances in lecture recognition: the ISL RT-06s evaluation system, in ‘Proceedings of Interspeech’, ISCA, Pittsburgh, Pennsylvania.
- Fukuhara, T., Nakagawa, H. and Nishida, T. (2007), Understanding Sentiment of People from News Articles: Temporal Sentiment Analysis of Social Events, in ‘International Conference on Web and Social Media (ICWSM)’.
- Gales, M. and Young, S. (2007), *The Application of Hidden Markov Models in Speech Recognition*, Vol. 1, Now Publishers Inc., Hanover, MA, USA.
- Galley, M., McKeown, K., Fosler-Lussier, E. and Jing, H. (2003), Discourse segmentation of multi-party conversation, in ‘Proceedings of the 41st Annual Meeting on Association for Computational Linguistics’, Association for Computational Linguistics, pp. 562–569.
- Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. (2005), Pulse: Mining customer opinions from free text, in ‘international symposium on intelligent data analysis’, Springer, pp. 121–132.

- Garcia-Fidalgo, E. and Ortiz, A. (2015), ‘Vision-based topological mapping and localization by means of local invariant features and map refinement’, *Robotics and Autonomous Systems* **64**, 1–25.
- Garsoffky, B., Schwan, S. and Huff, M. (2009), ‘Canonical views of dynamic scenes’, *Journal of Experimental Psychology: Human Perception and Performance* **35**(1), 17.
- Georgescu, M., Clark, A. and Armstrong, S. (2006a), An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms, in ‘Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue’, Association for Computational Linguistics, pp. 144–151.
- Georgescu, M., Clark, A. and Armstrong, S. (2006b), Word distributions for thematic segmentation in a support vector machine approach, in ‘Proceedings of the Tenth Conference on Computational Natural Language Learning’, Association for Computational Linguistics, pp. 101–108.
- Georgescu, M., Clark, A. and Armstrong, S. (2007), Exploiting structural meeting-specific features for topic segmentation, in ‘Actes de TALN/RECITAL’, Toulouse, France, pp. 15–24.
- Gillick, L. and Cox, S. J. (1989), Some statistical issues in the comparison of speech recognition algorithms, in ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 532–535.
- Glass, J. R., Hazen, T. J., Cyphers, D. S., Malioutov, I., Huynh, D. and Barzilay, R. (2007), Recent progress in the MIT spoken lecture processing project, in ‘Proceedings of Interspeech’, ISCA, Antwerp, Belgium, pp. 2553–2556.
- Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992), SWITCHBOARD: telephone speech corpus for research and development, in ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, Vol. 1, San Francisco, CA, pp. 517–520.
- Goedemé, T. and Van Gool, L. (2008), Robust vision-only mobile robot navigation with topological maps, in X.-J. Jing, ed., ‘Mobile Robots Mo-

- tion Planning, New Challenges', INTECH Open Access Publisher, Austria, chapter 4, pp. 63–88.
- Gokhan Tür and Hakkani-Tür, D. (2011), Human/Human Conversation Understanding, *in* G. Tür and R. De Mori, eds, 'Spoken Language Understanding: Systems for Extracting Semantic Information from Speech', chapter 9, pp. 227–255.
- Gopinath, R. A. (1998), Maximum likelihood modeling with Gaussian distributions for classification, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)', Vol. 2, pp. 661–664.
- Gorin, A. L., Parker, B. A., Sachs, R. M. and Wilpon, J. G. (1996), How may I help you?, *in* 'Proceedings of Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications', pp. 57–60.
- Griffiths, T. L. and Steyvers, M. (2002), A probabilistic approach to semantic representation, *in* 'Proceedings of the 24th annual conference of the cognitive science society', Citeseer, pp. 381–386.
- Griffiths, T. L. and Steyvers, M. (2004), 'Finding scientific topics', *Proceedings of the National Academy of Sciences of the United States of America* **101**(suppl 1), 5228–5235.
- Grosz, B. J. and Sidner, C. L. (1986), 'Attention, Intentions and the Structure of Discourse', *Computation Linguistics* **12**(3), 175–204.
- Haffner, P., Tür, G. and Wright, J. H. (2003), Optimizing SVMs for complex call classification, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Vol. 1, pp. 632–635.
- Hain, T., Burget, L., Dines, J., Garner, P. N., Grezl, F., Hannani, A. E., Huijbregts, M., Karafiat, M., Lincoln, M. and Wan, V. (2012), 'Transcribing Meetings With the AMIDA Systems', *IEEE Transactions on Audio, Speech, and Language Processing* **20**(2), 486–498.
- Hain, T., El Hannani, A., Wrigley, S. N. and Wan, V. (2008), Automatic speech recognition for scientific purposes - WebASR, *in* 'Proceedings of Interspeech', ISCA, Brisbane, Australia, pp. 504–507.

- Hakkani-Tür, D., Béchet, F., Riccardi, G. and Tür, G. (2006), ‘Beyond ASR 1-best: Using word confusion networks in spoken language understanding’, *Computer Speech and Language* **20**(4), 495–514.
- Halpern, Y., Hall, K., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G. and Baeuml, M. (2016), Contextual prediction models for speech recognition, *in* ‘Proceedings of Interspeech’, ISCA, Beijing, China.
- Han, J., Pei, J. and Kamber, M. (2011), Measuring data similarity and dissimilarity, *in* ‘Data mining: concepts and techniques’, 3 edn, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, chapter 2.4.
- Hazen, T. J. (2011), Topic Identification, *in* G. Tür and R. De Mori, eds, ‘Spoken Language Understanding: Systems for Extracting Semantic Information from Speech’, chapter 12, pp. 319–356.
- Hazen, T. J. and Richardson, F. (2008), A hybrid SVM/MCE training approach for vector space topic identification of spoken audio recordings, *in* ‘Proceedings of Interspeech’, ISCA, Brisbane, Australia, pp. 2542–2545.
- Hearst, M. a. (1997), ‘TextTiling: Segmenting text into multi-paragraph subtopic passages’, *Computational Linguistics* **23**(1), 33–64.
- Hearst, M. A. and Plaunt, C. (1993), Subtopic structuring for full-length document access, *in* ‘Proceedings of the international ACM SIGIR conference on Research and development in information retrieval’, ACM, pp. 59–68.
- Hemphill, C. T., Godfrey, J. J. and Doddington, G. R. (1990), ‘The ATIS Spoken Language Systems Pilot Corpus’, *Proceedings of the DARPA Speech and Natural Language Workshop* pp. 96–101.
- Hermansky, H. (1990), ‘Perceptual linear predictive (PLP) analysis of speech’, *the Journal of the Acoustical Society of America* **87**(4), 1738–1752.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. and Others (2012), ‘Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups’, *IEEE Signal Processing Magazine* **29**(6), 82–97.

- Hirschberg, J. and Litman, D. (1993), ‘Empirical studies on the disambiguation of cue phrases’, *Computational linguistics* **19**(3), 501–530.
- Hirschberg, J. and Nakatani, C. (1998), Acoustic Indicators of Topic Segmentation, *in* ‘Proceedings of International Conference on Spoken Language Processing (ICSLP)’.
- Hofmann, T. (1999), Probabilistic Latent Semantic Indexing, *in* ‘Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval’, ACM, New York, NY, USA, pp. 50–57.
- Holmes, J. and Holmes, W. (2001), *Speech synthesis and recognition*, Vol. 2, 2nd edn, Taylor & Francis, London.
- Hori, T., Hori, C., Minami, Y. and Nakamura, A. (2007), ‘Efficient WFST-Based One-Pass Decoding With On-The-Fly Hypothesis Rescoring in Extremely Large Vocabulary Continuous Speech Recognition’, *IEEE Transactions on Audio, Speech, and Language Processing* **15**(4), 1352–1365.
- Hsueh, P.-Y., Moore, J. D. and Renals, S. (2006), Automatic Segmentation of Multiparty Dialogue, *in* ‘the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)’.
- Hu, D. and Saul, L. K. (2009), A Probabilistic Topic Model for Unsupervised Learning of Musical Key-Profiles, *in* ‘Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)’, Citeseer, Kobe, Japan, pp. 441–446.
- Iyer, R. M. and Ostendorf, M. (1999), ‘Modeling long distance dependence in language: Topic mixtures versus dynamic cache models’, *IEEE Transactions on speech and audio processing* **7**(1), 30–39.
- Jelinek, F., Mercer, R. L., Bahl, L. R. and Baker, J. K. (1977), ‘Perplexity—a measure of the difficulty of speech recognition tasks’, *The Journal of the Acoustical Society of America* **62**(S1), S63–S63.
- Jones, K. S. (1972), ‘A statistical interpretation of term specificity and its application in retrieval’, *Journal of documentation* **28**(1), 11–21.



- Jurafsky, D. and Martin, J. H. (2009), *Speech And Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edn, Pearson, New Jersey.
- Kalashnikov, D. V., Hakkani-Tür, D., Tür, G. and Venkatasubramanian, N. (2009), Speech-Based Situational Awareness for Crisis Response, *in* 'EMWS DHS Workshop'.
- Kim, S., Narayanan, S. and Sundaram, S. (2009), Acoustic topic model for audio information retrieval, *in* 'IEEE Workshop on Applications of Signal Processing to Audio and Acoustics', pp. 37–40.
- Kim, W., Khudanpur, S. and Wu, J. (2001), Smoothing issues in the structured language model, *in* 'Proceedings of Interspeech', ISCA, pp. 717–720.
- Kneser, R. and Ney, H. (1995), Improved backing-off for m-gram language modeling, *in* 'IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Vol. 1, IEEE, pp. 181–184.
- Kuipers, B. (1978), 'Modeling spatial knowledge', *Cognitive Science* **2**(2), 129–153.
- Kuipers, B. and Byun, Y.-T. (1991), 'A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations', *Robotics and autonomous systems* **8**, 47–63.
- Kullback, S. and Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.
- Kuo, H. K. J. and Lee, C.-H. (2003), 'Discriminative training of natural language call routers', *IEEE Transactions on Speech and Audio Processing* **11**(1), 24–35.
- Kushner, W. M., Harton, S. M., Novorita, R. J. and McLaughlin, M. J. (2006), 'The acoustic properties of SCBA equipment and its effects on speech communication', *IEEE Communications Magazine* **44**(1), 66–72.
- Li, J., Tsao, Y. and Lee, C.-H. (2005), A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition, *in*

- ‘IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, pp. 837–840.
- Li, Y., Erdogan, H., Gao, Y. and Marcheret, E. (2002), Incremental on-line feature space MLLR adaptation for telephony speech recognition, *in* ‘Proceedings of Interspeech’, ISCA, Denver, Colorado, USA, pp. 1417–1420.
- Liberman, C. B., Geoffrois, E., Wu, Z. and Mark (1998), Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech, *in* ‘First International Conference on Language Resources and Evaluation (LREC)’, pp. 1373–1376.  
**URL:** <http://trans.sourceforge.net>
- Lisowska, A. (2003), Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study, Technical report, IM2.MDM-11, ISSCO, University of Geneva.
- Liu, M. and Siegwart, R. (2013), ‘Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera’, *IEEE Transactions on Robotics* **30**(2), 310–324.
- Löffler, J., Schon, J. and Köhler, J. (2006), SHARE : Supporting Large-Scale Rescue Operations with Communication and Information Services over Mobile Networks, *in* ‘2nd international conference on Mobile multimedia communications’, ACM, p. 47.
- Lombard, E. (1911), ‘Le signe de l’elevation de la voix’, *Ann. Maladies Oreille, Larynx, Nez, Pharynx* **37**(101-119), 25.
- Lucas-Cuesta, J. M., Ferreiros, J., Fernández-Martínez, F., Echeverry, J. D. and Lutfi, S. (2013), ‘On the dynamic adaptation of language models based on dialogue information’, *Expert Systems with Applications* **40**(4), 1069–1085.
- Lynch, K. (1960), *The image of the city*, Vol. 11, The MIT press.
- Makhoul, J., Kubala, F., Leek, T., Liu, D., Nguyen, L., Schwartz, R. and Srivastava, A. (2000), ‘Speech and language technologies for audio indexing and retrieval’, *Proceedings of the IEEE* **88**(8), 1338–1353.

- Malioutov, I. and Barzilay, R. (2006), Minimum cut model for spoken lecture segmentation, *in* ‘Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics’, Association for Computational Linguistics, pp. 25–32.
- Manning, C. D., Raghavan, P. and Schütze, H. (2009), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England.
- Martinez, F., Tapias, D. and Alvarez, J. (1998), Towards speech rate independence in large vocabulary continuous speech recognition, *in* ‘Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, Vol. 2, pp. 725–728.
- Massie, T. and Wijesekera, D. (2008), TVIS: Tactical voice interaction services for dismounted Urban operations, *in* ‘Proceedings of IEEE Military Communications Conference (MILCOM)’, pp. 1–7.
- Maybury, M. T. (1998), Discourse cues for broadcast news segmentation, *in* ‘Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics’, Association for Computational Linguistics, pp. 819–822.
- Mehrotra, S., Butts, C., Kalashnikov, D. V., Venkatasubramanian, N., Altintas, K., Hariharan, R., Lee, H., Ma, Y., Myers, A. and Wickramasuriya, J. (2004), CAMAS: a citizen awareness system for crisis mitigation, *in* ‘Proceedings of the 2004 ACM SIGMOD international conference on Management of data’, SIGMOD ’04, ACM, NY, USA, pp. 955–956.
- Mikolov, T. and Zweig, G. (2012), Context dependent recurrent neural network language model, *in* ‘IEEE Workshop on Spoken Language Technology (SLT)’, FL, USA, pp. 234–239.
- Miller, D., Boisen, S., Schwartz, R., Stone, R. and Weischedel, R. (2000), Named entity extraction from noisy input: speech and OCR, *in* ‘Proceedings of the sixth conference on Applied natural language processing’, Association for Computational Linguistics, pp. 316–324.

- Moore, J., Kronenthal, M. and Ashby, S. (2016), ‘AMI transcription’.  
**URL:** <http://groups.inf.ed.ac.uk/ami/corpus/transcription.shtml>
- Moore, R. K. (2016), ‘Introducing a pictographic language for envisioning a rich variety of enactive systems with different degrees of complexity’, *International Journal of Advanced Robotic Systems* **13**(2), 74.
- Morchid, M., Dufour, R., Bousquet, P. M., Bouallegue, M., Linares, G. and Mori, R. D. (2014a), Improving dialogue classification using a topic space representation and a Gaussian classifier based on the decision rule, *in* ‘IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, pp. 126–130.
- Morchid, M., Dufour, R. and Linares, G. (2014b), A LDA-based Topic Classification Approach from highly Imperfect Automatic Transcriptions, *in* ‘the 9th edition of the Language Resources and Evaluation Conference (LREC)’, Reykjavik, Iceland, pp. 1309–1314.
- Morgan, W., Chang, P.-C., Gupta, S. and Brenier, J. M. (2009), Automatically detecting action items in audio meeting recordings, *in* ‘Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue’, Association for Computational Linguistics, pp. 96–103.
- MOTU-896Mk3 (2016), ‘An audio interface/mixer’.  
**URL:** <http://www.motu.com/products/motuaudio/896mk3>
- Mowshowitz, A. and Dehmer, M. (2012), ‘Entropy and the complexity of graphs revisited’, *Entropy* **14**(3), 559–570.
- Murveit, H., Butzberger, J., Digalakis, V. and Weintraub, M. (1993), Large-vocabulary dictation using SRI’s DECIPHER speech recognition system: progressive search techniques, *in* ‘IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, Vol. 2, pp. 319–322.
- Nanjo, H. and Kawahara, T. (2003), Unsupervised language model adaptation for lecture speech recognition, *in* ‘ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition’.

- Nanjo, H. and Kawahara, T. (2004), ‘Language model and speaking rate adaptation for spontaneous presentation speech recognition’, *IEEE Transactions on Speech and Audio Processing* **12**(4), 391–400.
- NFPA (2014), Communications and Information Management, in ‘Standard on Emergency Services Incident Management System and Command Safety’, National Fire Protection Association (NFPA) 1561, chapter 6.
- NIST (2016), ‘Speech Recognition Scoring Toolkit (SCTK)’.  
**URL:** <https://www.nist.gov/itl/iad/mig/tools>
- NYS-USAR (2007), New York State Urban Search and Rescue Response System, Technical report, New York State Office of Fire Prevention and Control, Homeland Security and Emergency Services.
- Olney, A. and Cai, Z. (2005), An orthonormal basis for topic segmentation in tutorial dialogue, in ‘Proceedings of the conference on human language technology and empirical methods in natural language processing’, Association for Computational Linguistics, pp. 971–978.
- Pallotta, V., Niekrasz, J. and Purver, M. (2005), Collaborative and argumentative models of meeting discussions, in ‘Proceeding of CMNA-05 international workshop on Computational Models of Natural Arguments’.
- Palmer, D. D. (1999), Robust information extraction from spoken language data, in ‘Proceedings of Eurospeech’, ISCA, Budapest, Hungary.
- Passonneau, R. J. and Litman, D. J. (1997), ‘Discourse Segmentation by Human and Automated Means’, *Computational Linguistics* **23**(1), 103–139.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), ‘Scikit-learn: Machine Learning in Python’, *Journal of Machine Learning Research* **12**, 2825–2830.
- Pevzner, L. and Hearst, M. a. (2002), ‘A Critique and Improvement of an Evaluation Metric for Text Segmentation’, *Computational Linguistics* **28**(1), 19–36.

- Piaget, J. and Inhelder, B. (1956), *The Child's Conception of Space*, English translated by Langdon, F. J. and Lunzer, J. L., First published 1948, Routledge, London, UK.
- Pigeon, S., Swail, C., Geoffrois, E., Bruckner, C., van Leeuwen, D., Teixeira, C., Orman, O., Collins, P., Anderson, T., Grieco, J. and Zissman, M. (2005), Use of Speech and Language Technology in Military Environments, Technical Report ROT-TR-IST-037, Research and Technology Organisation of North Atlantic Treaty Organisation (NATO), Cedex, France.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K. (2011), The kaldi speech recognition toolkit, *in* 'IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)'.
- Przybocki, M. A., Fiscus, J. G., Garofolo, J. S. and Pallett, D. S. (1999), 1998 Hub-4 information extraction evaluation, *in* 'Proceedings of DARPA Broadcast News Workshop', Va, USA, pp. 13–18.
- Purver, M. (2011), Topic Segmentation, *in* G. Tür and R. De Mori, eds, 'Spoken Language Understanding: Systems for Extracting Semantic Information from Speech', chapter 11, pp. 291–317.
- Purver, M., Griffiths, T. L., Körding, K. P. and Tenenbaum, J. B. (2006), Un-supervised topic modelling for multi-party spoken discourse, *in* 'Proceedings of International Conference on Computational Linguistics', Association for Computational Linguistics, pp. 17–24.
- Rabiner, L. L. and Juang, B.-H. B. (1993), *Fundamentals of Speech Recognition*, Vol. 103, Prentice Hall International.
- Ranganathan, A. and Dellaert, F. (2011), 'Online probabilistic topological mapping', *The International Journal of Robotics Research* **30**(6), 755–771.
- Rehurek, R. and Sojka, P. (2010), Software framework for topic modelling with large corpora, *in* 'Proceedings of the International Conference on Language Resources and Evaluation (LREC)', Workshop on New Challenges for NLP Frameworks', ELRA, Valletta, Malta, pp. 45–50.

- Remolina, E. and Kuipers, B. (2004), ‘Towards a general theory of topological maps’, *Artificial Intelligence* **152**(1), 47–104.
- Reynar, J. C. (1994), An automatic method of finding topic boundaries, *in* ‘Proceedings of the 32nd annual meeting on Association for Computational Linguistics’, Association for Computational Linguistics, pp. 331–333.
- Reynar, J. C. (1998), Topic segmentation: Algorithms and applications, PhD thesis, University of Pennsylvania.
- Robinson, A. (1996), ‘The British English Example Pronunciation (BEEP) dictionary’.  
**URL:** <http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html>
- Rose, R. C., Chang, E. I. and Lippmann, R. P. (1991), Techniques for information retrieval from voice messages, *in* ‘Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, pp. 317–320 vol.1.
- Rosenfeld, R. (1995), Optimizing lexical and ngram coverage via judicious use of linguistic data, *in* ‘Proceedings of Eurospeech’, ISCA, Madrid, Spain, pp. 1763–1766.
- Rosenfeld, R. (2000), ‘Two decades of statistical language modeling: Where do we go from here?’, *Proceedings of the IEEE* **88**(8).
- Sak, H., Saraçlar, M. and Güngör, T. (2010), On-the-fly lattice rescoring for real-time automatic speech recognition, *in* ‘Proceedings of Interspeech’, ISCA, pp. 2450–2453.
- Schaitberger, H., Miller, T. and Morrison, P. (2016), *Voice Radio Communications Guide for the Fire Service*, june edn, Federal Emergency Management Agency (FEMA).
- Schneider, D., Winkler, T., Löffler, J. and Schon, J. (2007), Robust audio indexing and keyword retrieval optimized for the rescue operation domain, *in* ‘International Workshop on Mobile Information Technology for Emergency Response’, Springer Berlin Heidelberg, pp. 135–142.

- Schwartz, R. and Chow, Y. L. (1990), The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses, *in* ‘IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)’, pp. 81–84 vol.1.
- Sebastiani, F. (2002), ‘Machine learning in automated text categorization’, *ACM computing surveys (CSUR)* **34**(1), 1–47.
- Seide, F., Li, G. and Yu, D. (2011), Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, *in* ‘Proceedings of Interspeech’, ISCA, pp. 437–440.
- Seymore, K., Chen, S. F. and Rosenfeld, R. (1998), Nonlinear interpolation of topic models for language model adaptation, *in* ‘Proceedings of International Conference on Spoken Language Processing (ICSLP)’, Sydney, Australia.
- Seymore, K. and Rosenfeld, R. (1997), Using story topics for language model adaptation, Technical report, Carnegie Mellon University.
- Shi, Y. (2014), Language models with meta-information, Phd thesis, TU Delft, Delft University of Technology.
- Shimanski, C. (2008), Situational Awareness in Search and Rescue Operations, Technical report, Mount Hood, Oregon, USA.
- Shriberg, E. (1996), Disfluencies in switchboard, *in* ‘Proceedings of International Conference on Spoken Language Processing (ICSLP)’, Vol. 96, pp. 11–14.
- Siegel, A. W. and White, S. H. (1975), ‘The Development of Spatial Representations of Large-Scale Environments’, *Advances in Child Development and Behavior* **10**, 9–55.
- Siniscalchi, S. M., Li, J. and Lee, C. H. (2006), A study on lattice rescoring with knowledge scores for automatic speech recognition, *in* ‘Proceedings of Interspeech’, pp. 517–520.
- Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A. and Freeman, W. T. (2005), Discovering objects and their location in images, *in* ‘10th IEEE



- International Conference on Computer Vision (ICCV)', Vol. 1, Beijing, Chin, pp. 370–377.
- Stankiewicz, B. J., Legge, G. E., Mansfield, J. S. and Schlicht, E. J. (2006), 'Lost in virtual space: studies in human and ideal spatial navigation', *Journal of Experimental Psychology: Human Perception and Performance* **32**(3), 688.
- Stein, D., Schwenninger, J. and Usabaev, B. (2012), Automatic Chat Transcription on a Firefighter TETRA Broadcast Channel, in 'Proceedings of Speech Communication; 10. ITG Symposium', VDE, pp. 1–4.
- Stein, D. and Usabaev, B. (2012), Automatic Speech Recognition on Firefighter TETRA broadcast, in 'Proceedings of the International Conference on Language Resources and Evaluation (LREC)', European Language Resources Association (ELRA), Istanbul, Turkey.
- Stent, A. J. (2001), The Monroe Corpus, Technical report, University of Rochester, Rochester, NY, USA.
- Stolcke, A. (2002), Srlm-an Extensible Language Modeling Toolkit, in 'Proceedings of Interspeech', ISCA, Colorado, USA, pp. 901–904.
- Stupakov, A., Hanusa, E., Vijaywargi, D., Fox, D. and Bilmes, J. (2012), 'The design and collection of COSINE, a multi-microphone in situ speech corpus recorded in noisy environments', *Computer Speech and Language* **26**(1), 52–66.
- Tang, M., Pellom, B. and Hacioglu, K. (2003), Call-type classification and unsupervised training for the call center domain, in 'IEEE Workshop on Automatic Speech Recognition and Understanding', pp. 204–208.
- Tapus, A. and Siegwart, R. (2005), 'Incremental robot mapping with fingerprints of places', *IEEE International Conference on Intelligent Robots and Systems (IROS)* pp. 172–177.
- Thrun, S. (2002), Robotic Mapping: A Survey, in G. Lakemeyer and B. Nebel, eds, 'Exploring Artificial Intelligence in the New Millennium', Elsevier science (USA), chapter 1, pp. 1–35.

- Trancoso, I., Nunes, R., Neves, L., Viana, C., Moniz, H., Caseiro, D. and Mata, A. I. (2006), Recognition of classroom lectures in european portuguese, *in* ‘Proceedings of Interspeech’, ISCA, Pittsburgh, USA.
- Tür, G. and De Mori, R. (2011), *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, John Wiley & Sons.
- Tür, G., Hakkani-Tür, D., Stolcke, A. and Shriberg, E. (2001), ‘Integrating prosodic and lexical cues for automatic topic segmentation’, *Computational linguistics* **27**(1), 31–57.
- UN-OCHA (2012), The International Search and Rescue Advisory Group Guidelines and Methodology, Technical report, United Nations Office For The Coordination Of Humanitarian Affairs, Field Coordination Support Section (INSARAG Secretariat).
- URL:** <http://www.insarag.org>
- Unity (2016), ‘Unity (Personal): a 3D game engine development platform.’
- URL:** <https://unity3d.com/>
- Utiyama, M. and Isahara, H. (2001), A statistical model for domain-independent text segmentation, *in* ‘Proceedings of the 39th Annual Meeting on Association for Computational Linguistics’, ACL, pp. 499–506.
- Viterbi, A. (1967), ‘Error bounds for convolutional codes and an asymptotically optimum decoding algorithm’, *IEEE Transactions on Information Theory* **13**(2), 260–269.
- Walker, P. (1991), ‘International search and rescue teams : A League discussion paper’, *League of Red Cross and Red Crescent Societies* **1**(1375).
- Wei, X. and Croft, W. B. (2006), LDA-based document models for ad-hoc retrieval, *in* ‘Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval’, ACM, Seattle, Washington, USA, pp. 178–185.
- Wen, T.-H., Heidel, A., Lee, H.-Y., Tsao, Y. and Lee, L.-S. (2013), Recurrent neural network based personalized language modeling by social network crowdsourcing, *in* ‘Proceedings of Interspeech’, ISCA, Lyon, France, pp. 435–439.

- Weston, J. and Watkins, C. (1998), Multi-class support vector machines, Technical report, CSD-TR-98-04, Department of Computer Science, University of London.
- Wong, J. and Robinson, C. (2004), Urban search and rescue technology needs: identification of needs, Technical report.
- Wrede, B. and Shriberg, E. (2003), Relationship between dialogue acts and hot spots in meetings, *in* ‘IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)’, pp. 180–185.
- Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D. and Zweig, G. (2016), Achieving Human Parity in Conversational Speech Recognition, Technical report, Microsoft Research.
- Yamron, J. P., Carp, I., Gillick, L., Lowe, S. and van Mulbregt, P. (1998), A hidden Markov model approach to text segmentation and event tracking, *in* ‘Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)’, Vol. 1, pp. 333–336.
- Youmans, G. (1991), ‘A New Tool for Discourse Analysis: The Vocabulary Management Profile’, *Language* **67**, 763–789.
- Yu, D. and Seltzer, M. (2011), Improved Bottleneck Features Using Pretrained Deep Neural Networks, *in* ‘Proceedings of Interspeech’, ISCA, Florence, Italy.
- Yuan, G. X., Ho, C. H. and Lin, C. J. (2012), ‘Recent Advances of Large-Scale Linear Classification’, *Proceedings of the IEEE* **100**(9), 2584–2603.



# Appendix A

## Examples of TF-IDF

---

This appendix presents a list of first 20 trigrams received highest TF-IDF on a document of transition-related utterances and 13 documents of room-specific utterances.

**Transition:** "okay", "okay so", "room", "so", "okay in", "am in", "erm", "in the", "okay so in", "in er", "so in", "in room", "like", "first room", "started in", "okay i'am in", "okay i'am", "started in", "into the", "got"

**RoomType 1:** "two", "computer", "room", "desks", "computer room", "whiteboard", "chairs", "through", "that", "two computers", "door", "printer", "office", "doors", "the computer", "each", "then", "the computer room", "two desks", "each"

**RoomType 2:** "classroom", "like", "one", "blackboard", "desk", "the classroom", "the back", "desks", "teacher desk", "teacher", "room", "nine", "at the back", "through", "door", "back", "then there", "coat", "desk and", "classroom erm"

**RoomType 3:** "bathroom", "washing", "washing machine", "toilet", "machine", "room", "bath", "sink", "mirror", "one", "clothes", "the bathroom", "into", "hoover", "through", "rack", "door", "laundry", "doors", "radiator"

**RoomType 4:** "chest", "some", "the floor", "floor", "on the floor", "of drawers", "desk", "chest of drawers", "books", "chest of", "bed", "drawers", "poster", "like", "toys", "wall", "on the wall", "bedroom", "lamp", "child"

**RoomType 5:** "kitchen", "fridge", "sink", "hob", "oven", "room", "two", "microwave", "then", "the kitchen", "and then", "door", "in the", "cupboards", "another", "which", "dining", "through", "on the", "like"

**RoomType 6:** "vending", "vending machines", "machines", "tables", "man", "room", "there man", "vending machines and", "machines and", "three vending", "three vending machines", "door", "canteen", "there radiator", "radiator", "cafeteria", "on the floor", "the floor", "got", "floor"

**RoomType 7:** "table", "room", "gym", "games room", "games", "trampoline", "weights", "treadmills", "machine", "door", "machines", "exercise", "table tennis", "pool", "cycling", "treadmill", "mats", "tennis table", "pool table", "cross"

**RoomType 8:** "bed", "bedroom", "room", "the bed", "double", "double bed", "bedside", "lamp", "which", "door", "radiator", "table", "on the", "through", "red", "into", "got", "wardrobe", "of the", "bed with"

**RoomType 9:** "room", "living", "living room", "one", "into", "through", "sofas", "grandfather", "the living", "grandfather clock", "rug", "the living room", "clock", "fan", "back", "left", "into the", "doors", "ceiling", "door"

**RoomType 10:** "desks", "bookshelves", "books", "library", "shelves", "like", "study", "bookshelf", "three desks", "sort of", "sort", "and two", "smaller", "big", "back", "this room", "books and", "desks er", "desks and", "going"

**RoomType 11:** "dining", "dining room", "the dining room", "the dining", "chairs", "four", "back", "clock", "through", "go", "table", "doors", "tables", "which", "going", "on the", "door", "wall", "two", "clock on"

**RoomType 12:** "like", "fire", "some", "the fire", "work", "sort", "erm there", "sort of", "extinguisher", "drill", "looks", "workshop", "fire extinguisher", "tools", "kind", "blue", "kind of", "saw", "shelves", "looks like"

**RoomType 13:** "boiler", "pipes", "boiler room", "like", "two boilers", "the wall", "this room", "sure", "working", "not sure", "yeah", "back", "wall", "out of", "go", "on the", "of them", "on the wall", "no other", "erm so"

# Appendix B

## Examples of LDA

---

This appendix presents presents a list of {word, probability} for the 20 most probable words in 40 topics learnt on the manual transcripts of the Switchboard telephone speech corpus.

**Topic01** {nice, 0.09}, {dollars, 0.058}, {working, 0.052}, {thousand, 0.032}, {hundred, 0.031}, {fifty, 0.028}, {paid, 0.025}, {name, 0.023}, {several, 0.022}, {california, 0.02}, {miles, 0.02}, {top, 0.02}, {absolutely, 0.019}, {hope, 0.018}, {except, 0.015}, {rest, 0.015}, {out, 0.013}, {single, 0.013}, {oil, 0.012}, {guy, 0.012}

**Topic02** {go, 0.193}, {work, 0.114}, {home, 0.079}, {out, 0.044}, {when, 0.03}, {time, 0.024}, {gets, 0.018}, {stay, 0.017}, {spend, 0.017}, {wanna, 0.015}, {back, 0.015}, {want, 0.014}, {until, 0.014}, {often, 0.013}, {after, 0.013}, {ahead, 0.012}, {kids, 0.012}, {into, 0.012}, {outside, 0.012}, {because, 0.011}

**Topic03** {two, 0.095}, {years, 0.093}, {three, 0.059}, {five, 0.051}, {four, 0.039}, {twenty, 0.037}, {ago, 0.034}, {couple, 0.031}, {six, 0.027}, {times, 0.02}, {half, 0.02}, {thirty, 0.02}, {eight, 0.018}, {hundred, 0.016}, {months, 0.016}, {eighty, 0.015}, {percent, 0.015}, {only, 0.014}, {major, 0.012}, {days, 0.012}

**Topic04** {going, 0.178}, {said, 0.073}, {find, 0.046}, {keep, 0.044}, {area, 0.037}, {same, 0.033}, {point, 0.03}, {dallas, 0.029}, {up, 0.026}, {american, 0.014}, {out, 0.013}, {time, 0.013}, {group, 0.011}, {over, 0.011}, {told, 0.009}, {because, 0.009}, {stop, 0.009}, {newspapers, 0.009}, {watching, 0.009}, {hey, 0.009}

**Topic05** {right, 0.537}, {now, 0.248}, {spent, 0.013}, {north, 0.012}, {student, 0.01}, {regular, 0.009}, {build, 0.007}, {road, 0.006}, {mountains, 0.006},

{defense, 0.005}, {colorado, 0.004}, {treat, 0.003}, {spot, 0.003}, {carolina, 0.003}, {illegal, 0.003}, {essentially, 0.003}, {ends, 0.003}, {definite, 0.002}, {seriously, 0.002}, {fee, 0.002}

**Topic06** {like, 0.415}, {something, 0.071}, {school, 0.055}, {feel, 0.04}, {seems, 0.037}, {high, 0.023}, {college, 0.021}, {kind, 0.02}, {stuff, 0.018}, {public, 0.013}, {schools, 0.012}, {pick, 0.011}, {kids, 0.011}, {takes, 0.01}, {because, 0.01}, {son, 0.009}, {out, 0.009}, {eat, 0.008}, {everyone, 0.007}, {cans, 0.006}

**Topic07** {kids, 0.121}, {school, 0.058}, {high, 0.057}, {usually, 0.045}, {know, 0.039}, {end, 0.036}, {month, 0.028}, {education, 0.026}, {schools, 0.023}, {insurance, 0.02}, {when, 0.019}, {health, 0.019}, {public, 0.019}, {up, 0.018}, {going, 0.017}, {having, 0.014}, {because, 0.014}, {side, 0.014}, {hear, 0.013}, {second, 0.013}

**Topic08** {anything, 0.106}, {everything, 0.075}, {news, 0.06}, {else, 0.054}, {exactly, 0.052}, {may, 0.041}, {budget, 0.029}, {unless, 0.018}, {something, 0.018}, {personal, 0.014}, {bill, 0.013}, {late, 0.013}, {support, 0.013}, {throw, 0.012}, {current, 0.011}, {amazing, 0.011}, {television, 0.011}, {magazine, 0.009}, {finding, 0.008}, {season, 0.008}

**Topic09** {his, 0.071}, {family, 0.062}, {why, 0.057}, {better, 0.055}, {texas, 0.042}, {yet, 0.03}, {definitely, 0.028}, {crime, 0.027}, {situation, 0.026}, {less, 0.025}, {making, 0.022}, {love, 0.019}, {against, 0.018}, {plan, 0.016}, {who's, 0.014}, {social, 0.01}, {florida, 0.01}, {rate, 0.01}, {costs, 0.01}, {degree, 0.008}

**Topic10** {probably, 0.156}, {either, 0.054}, {agree, 0.044}, {child, 0.04}, {small, 0.036}, {certainly, 0.033}, {best, 0.029}, {seven, 0.025}, {gun, 0.025}, {wrong, 0.02}, {sixty, 0.015}, {national, 0.014}, {morning, 0.013}, {nineteen, 0.013}, {bunch, 0.012}, {size, 0.011}, {boys, 0.01}, {crazy, 0.01}, {rid, 0.009}, {green, 0.008}

**Topic11** {really, 0.458}, {live, 0.05}, {again, 0.036}, {let, 0.025}, {because, 0.02}, {tried, 0.016}, {using, 0.014}, {book, 0.013}, {much, 0.013}, {liked, 0.012}, {store, 0.011}, {longer, 0.011}, {happy, 0.01}, {kid, 0.01}, {ready, 0.009}, {obviously, 0.008}, {surprised, 0.007}, {though, 0.007}, {noticed, 0.007}, {rain, 0.006}

**Topic12** {take, 0.085}, {money, 0.071}, {care, 0.061}, {many, 0.05}, {being, 0.046}, {pay, 0.04}, {buy, 0.03}, {able, 0.03}, {everybody, 0.027}, {place, 0.027}, {people, 0.022}, {health, 0.019}, {important, 0.018}, {set, 0.017}, {involved, 0.016}, {taking, 0.016}, {insurance, 0.015}, {cases, 0.014}, {start, 0.013}, {much, 0.012}

**Topic13** {need, 0.088}, {course, 0.082}, {made, 0.053}, {women, 0.027}, {control, 0.023}, {terms, 0.02}, {under, 0.02}, {law, 0.018}, {lives, 0.018}, {taken, 0.017}, {needs, 0.017}, {looks, 0.016}, {level, 0.015}, {teachers, 0.015}, {hot, 0.013}, {men, 0.012}, {fall, 0.011}, {lost, 0.011}, {opinion, 0.01}, {by, 0.008}



- Topic14** {lot, 0.142}, {some, 0.137}, {know, 0.107}, {things, 0.105}, {those, 0.057}, {people, 0.057}, {these, 0.043}, {different, 0.031}, {other, 0.018}, {kind, 0.014}, {stuff, 0.011}, {recently, 0.009}, {because, 0.009}, {show, 0.008}, {out, 0.007}, {reading, 0.006}, {looked, 0.006}, {learn, 0.006}, {spending, 0.006}, {around, 0.005}
- Topic15** {know, 0.479}, {say, 0.057}, {most, 0.034}, {always, 0.029}, {look, 0.025}, {hard, 0.022}, {because, 0.022}, {something, 0.022}, {people, 0.02}, {job, 0.019}, {whether, 0.013}, {looking, 0.013}, {time, 0.012}, {makes, 0.01}, {much, 0.009}, {make, 0.008}, {sense, 0.007}, {mind, 0.007}, {question, 0.006}, {whatever, 0.006}
- Topic16** {any, 0.156}, {play, 0.04}, {problems, 0.038}, {seen, 0.038}, {other, 0.034}, {few, 0.032}, {happen, 0.026}, {hand, 0.02}, {kinds, 0.018}, {without, 0.018}, {wonder, 0.017}, {particular, 0.016}, {married, 0.016}, {interest, 0.015}, {war, 0.015}, {society, 0.015}, {imagine, 0.014}, {figure, 0.013}, {lately, 0.012}, {crimes, 0.01}
- Topic17** {yeah, 0.946}, {fault, 0.001}, {trend, 0.001}, {dish, 0.001}, {joke, 0.001}, {annual, 0.0}, {marriage, 0.0}, {ratio, 0.0}, {tries, 0.0}, {gain, 0.0}, {importance, 0.0}, {mavericks, 0.0}, {mother-in-law, 0.0}, {noon, 0.0}, {golfer, 0.0}, {counter, 0.0}, {attract, 0.0}, {sticks, 0.0}, {inclined, 0.0}, {failure, 0.0}
- Topic18** {here, 0.119}, {down, 0.066}, {back, 0.06}, {up, 0.049}, {long, 0.046}, {come, 0.035}, {out, 0.032}, {when, 0.025}, {quite, 0.025}, {came, 0.021}, {time, 0.02}, {took, 0.02}, {comes, 0.018}, {coming, 0.016}, {understand, 0.016}, {dog, 0.015}, {works, 0.014}, {happened, 0.013}, {because, 0.012}, {difficult, 0.01}
- Topic19** {way, 0.14}, {new, 0.083}, {still, 0.08}, {since, 0.039}, {ones, 0.034}, {gone, 0.028}, {business, 0.025}, {which, 0.025}, {york, 0.017}, {check, 0.015}, {vote, 0.015}, {wants, 0.013}, {out, 0.012}, {cards, 0.011}, {nobody, 0.011}, {write, 0.01}, {price, 0.01}, {same, 0.01}, {kind, 0.008}, {asked, 0.007}
- Topic20** {problem, 0.083}, {having, 0.054}, {company, 0.048}, {end, 0.036}, {drug, 0.031}, {change, 0.03}, {month, 0.025}, {wear, 0.023}, {test, 0.021}, {testing, 0.018}, {instead, 0.015}, {general, 0.015}, {weather, 0.015}, {up, 0.014}, {later, 0.012}, {families, 0.012}, {medical, 0.012}, {trash, 0.011}, {dress, 0.011}, {white, 0.011}
- Topic21** {sure, 0.157}, {interesting, 0.066}, {read, 0.06}, {computer, 0.027}, {today, 0.023}, {running, 0.022}, {funny, 0.021}, {neat, 0.021}, {second, 0.018}, {cost, 0.017}, {ah, 0.016}, {kind, 0.016}, {starting, 0.014}, {driving, 0.014}, {books, 0.013}, {students, 0.013}, {hold, 0.012}, {beautiful, 0.011}, {union, 0.011}, {shows, 0.01}
- Topic22** {about, 0.325}, {guess, 0.151}, {bad, 0.04}, {enough, 0.037}, {talking, 0.029}, {talk, 0.027}, {world, 0.021}, {thinking, 0.02}, {supposed, 0.018},

{thought, 0.017}, {camping, 0.015}, {between, 0.013}, {weeks, 0.012}, {team, 0.008}, {talked, 0.008}, {concerned, 0.008}, {hours, 0.008}, {football, 0.006}, {much, 0.006}, {happening, 0.006}

**Topic23** {more, 0.176}, {little, 0.137}, {than, 0.094}, {maybe, 0.063}, {sort, 0.061}, {bit, 0.055}, {stuff, 0.031}, {paper, 0.024}, {hum, 0.018}, {much, 0.016}, {other, 0.013}, {kind, 0.011}, {quite, 0.01}, {fine, 0.009}, "won't, 0.009", 'areas, 0.009', {might, 0.008}, {used, 0.008}, {bigger, 0.006}, {available, 0.006}

**Topic24** {house, 0.071}, {great, 0.07}, {watch, 0.051}, {anyway, 0.045}, {fun, 0.044}, {both, 0.043}, {enjoy, 0.033}, {bought, 0.032}, {wonderful, 0.028}, {yourself, 0.022}, {fairly, 0.019}, {mine, 0.017}, {up, 0.017}, {somewhere, 0.016}, {t.v., 0.014}, {fish, 0.013}, {kind, 0.013}, {favorite, 0.012}, {built, 0.011}, {making, 0.011}

**Topic25** {mean, 0.299}, {okay, 0.187}, {believe, 0.033}, {large, 0.021}, {case, 0.02}, {expensive, 0.019}, {reason, 0.019}, {experience, 0.018}, {guns, 0.015}, {exercise, 0.015}, {local, 0.014}, {necessarily, 0.011}, {feeling, 0.01}, {short, 0.01}, {fair, 0.009}, {daughter, 0.009}, {bye-bye, 0.009}, {anyone, 0.008}, {individual, 0.008}, {bye, 0.007}

**Topic26** {her, 0.146}, {husband, 0.053}, {wanted, 0.046}, {help, 0.038}, {benefits, 0.036}, {young, 0.032}, {mother, 0.029}, {food, 0.026}, {service, 0.024}, {research, 0.017}, {because, 0.016}, {cause, 0.016}, {whenever, 0.015}, {killed, 0.013}, {restaurant, 0.009}, {kind, 0.009}, {children, 0.009}, {groups, 0.008}, {rights, 0.008}, {by, 0.007}

**Topic27** {oh, 0.528}, {see, 0.186}, {true, 0.062}, {boy, 0.021}, {gosh, 0.013}, {bet, 0.012}, {goodness, 0.01}, {god, 0.009}, {mexico, 0.006}, {planning, 0.005}, {common, 0.005}, {product, 0.004}, {man, 0.004}, {anytime, 0.003}, {limited, 0.003}, {poor, 0.003}, {round, 0.002}, {neighbor, 0.002}, {numbers, 0.002}, {present, 0.002}

**Topic28** {who, 0.089}, {use, 0.052}, {gonna, 0.044}, {try, 0.039}, {somebody, 0.035}, {state, 0.03}, {give, 0.025}, {especially, 0.024}, {people, 0.023}, {seem, 0.022}, {tax, 0.021}, {taxes, 0.021}, {someone, 0.019}, {amount, 0.016}, {nothing, 0.016}, {punishment, 0.014}, {capital, 0.014}, {paying, 0.014}, {certain, 0.014}, {pay, 0.013}

**Topic29** {huh, 0.083}, {fact, 0.064}, {usually, 0.059}, {sometimes, 0.058}, {tell, 0.053}, {matter, 0.027}, {nursing, 0.024}, {difference, 0.021}, {leave, 0.017}, {felt, 0.016}, {information, 0.015}, {homes, 0.014}, {consider, 0.013}, {choice, 0.013}, {street, 0.013}, {worry, 0.012}, {savings, 0.012}, {biggest, 0.012}, {cats, 0.011}, {unfortunately, 0.011}

**Topic30** {good, 0.184}, {very, 0.128}, {pretty, 0.103}, {real, 0.081}, {much, 0.043}, {life, 0.023}, {wow, 0.022}, {idea, 0.02}, {sounds, 0.02}, {program, 0.019}, {easy, 0.015}, {friend, 0.012}, {movie, 0.012}, {newspaper, 0.009}, {issue,

0.009}, {which, 0.007}, {buying, 0.007}, {caught, 0.007}, {realize, 0.006}, {woman, 0.006}

**Topic31** {what, 0.375}, {doing, 0.067}, {through, 0.037}, {kind, 0.03}, {system, 0.026}, {saying, 0.023}, {basically, 0.022}, {jury, 0.017}, {states, 0.012}, {education, 0.012}, {happens, 0.01}, {university, 0.009}, {by, 0.008}, {countries, 0.008}, {other, 0.008}, {needed, 0.007}, {completely, 0.006}, {decide, 0.006}, {united, 0.006}, {trial, 0.006}

**Topic32** {thing, 0.146}, {big, 0.069}, {one, 0.057}, {only, 0.055}, {kind, 0.045}, {another, 0.042}, {type, 0.036}, {country, 0.03}, {city, 0.022}, {music, 0.021}, {such, 0.02}, {wife, 0.017}, {deal, 0.016}, {hear, 0.016}, {listen, 0.015}, {same, 0.015}, {radio, 0.01}, {other, 0.009}, {teacher, 0.008}, {stuff, 0.007}

**Topic33** {getting, 0.077}, {far, 0.056}, {away, 0.044}, {run, 0.031}, {cars, 0.028}, {together, 0.028}, {myself, 0.028}, {gotten, 0.022}, {drive, 0.022}, {close, 0.02}, {left, 0.019}, {older, 0.019}, {hour, 0.017}, {tend, 0.017}, {lots, 0.016}, {into, 0.016}, {community, 0.014}, {by, 0.013}, {become, 0.013}, {ways, 0.012}

**Topic34** {get, 0.265}, {every, 0.051}, {car, 0.048}, {into, 0.038}, {trying, 0.034}, {once, 0.03}, {out, 0.03}, {while, 0.028}, {when, 0.024}, {up, 0.017}, {anymore, 0.016}, {time, 0.014}, {because, 0.012}, {worth, 0.011}, {kind, 0.01}, {cold, 0.009}, {much, 0.008}, {phone, 0.008}, {wish, 0.007}, {extra, 0.007}

**Topic35** {they, 0.523}, {know, 0.099}, {their, 0.08}, {want, 0.044}, {because, 0.028}, {people, 0.019}, {couldn't, 0.012}, {out, 0.01}, {up, 0.007}, {themselves, 0.006}, {call, 0.006}, {make, 0.006}, {come, 0.005}, {whatever, 0.004}, {quality, 0.004}, {gave, 0.004}, {glass, 0.004}, {kill, 0.003}, {start, 0.003}, {carry, 0.002}

**Topic36** {think, 0.475}, {should, 0.054}, {government, 0.027}, {parents, 0.022}, {found, 0.019}, {people, 0.017}, {person, 0.013}, {anybody, 0.013}, {office, 0.01}, {bring, 0.01}, {magazines, 0.009}, {because, 0.009}, {ought, 0.009}, {totally, 0.008}, {plastic, 0.007}, {seemed, 0.006}, {inside, 0.006}, {much, 0.006}, {bottles, 0.006}, {garbage, 0.005}

**Topic37** {one, 0.214}, {year, 0.072}, {time, 0.055}, {day, 0.052}, {kids, 0.044}, {old, 0.042}, {last, 0.034}, {ten, 0.026}, {remember, 0.021}, {week, 0.02}, {next, 0.019}, {night, 0.017}, {saw, 0.016}, {other, 0.014}, {summer, 0.012}, {during, 0.011}, {interested, 0.01}, {twelve, 0.008}, {christmas, 0.008}, {seventy, 0.007}

**Topic38** {your, 0.193}, {where, 0.19}, {own, 0.046}, {number, 0.026}, {call, 0.021}, {says, 0.018}, {death, 0.017}, {sit, 0.016}, {down, 0.015}, {cut, 0.015}, {trouble, 0.013}, {up, 0.013}, {game, 0.012}, {whatever, 0.011}, {utah, 0.009}, {penalty, 0.008}, {yard, 0.007}, {out, 0.007}, {environment, 0.007}, {salary, 0.007}

**Topic39** {um-hum, 0.774}, {air, 0.016}, {federal, 0.011}, {effect, 0.008}, {would've, 0.006}, {deterrent, 0.005}, {warm, 0.004}, {pop, 0.004}, {aerobics, 0.004}, {pickup, 0.004}, {focus, 0.004}, {adult, 0.004}, {chicago, 0.003}, {excited, 0.003}, {outrageous, 0.002}, {element, 0.002}, {conditioning, 0.002}, {dying, 0.002}, {awfully, 0.002}, {caused, 0.002}

**Topic40** {done, 0.101}, {although, 0.032}, {along, 0.025}, {water, 0.025}, {up, 0.022}, {line, 0.022}, {cat, 0.02}, {space, 0.019}, {dad, 0.018}, {move, 0.015}, {played, 0.014}, {south, 0.014}, {around, 0.014}, {story, 0.014}, {mom, 0.013}, {apparently, 0.012}, {aids, 0.012}, {higher, 0.011}, {policy, 0.01}, {towards, 0.01}

## SSAR Corpus Recording Forms

---

This appendix presents the SSAR corpus recording forms. The following documents were provided to each participant by email one day before the recording and also in printed form on the day of recording:

- The ‘*Information Sheet*’ explains the recording and instruct each participant how to perform the task.
- The ‘*Personal Information*’ form is used to collect participants personal information. All the provided information except the subjects' name and email address are available in the dataset.
- Each participant together with the lead researcher signed and dated two copies of the ‘*Consent Form*’ on the day of recording and received one copy for they own record.

## C.1 Information sheet

**The University of Sheffield**

### Information Sheet

## The Role of Voice Communications in the Search and Rescue Environment

---

### Researchers:

Lead researcher: Saeid Mokaram ([s.mokaram@sheffield.ac.uk](mailto:s.mokaram@sheffield.ac.uk))

Supervisor: Professor Roger Moore ([r.k.moore@sheffield.ac.uk](mailto:r.k.moore@sheffield.ac.uk))

### Invitation:

You are being invited to take part in a research project. Before you decide whether to participate or not, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or if you would like more information.

### Aim:

The purpose of this experiment is to investigate an automatic solution for extracting and structuring valuable information flowing on voice communication channels during crisis response among responders. For example, imagine the communications between the firefighters and command and control station. Firefighters explore the incident scene and communicate with the Task Force Leader (TFL) in the command and control station. The TFL tries to make a report of the situation by listening to the firefighter's report. We aim to design a system that can assist TFL by automatically extracting and structuring valuable information from voice communication channels.

### Your task:

In this experiment, you and another volunteer will be participating in a conversation which is designed to mimic a firefighter's communications to the command and control station. The firefighter explores a virtual environment (computer simulation) and communicates with the Task Force Leader (TFL) in the command and control station. The TFL tries to make a report of the situation by only listening to the firefighter's report.

The computer simulated environment is very similar to the conventional computer 3D games. A practice session will be given prior to the experiments to help participants familiarizing themselves how to move and control their avatar in the simulated environment.

The experiment will be conducted in a sound-attenuating booth for avoiding external acoustic disturbances. The sound-attenuating booth is located in the Speech and Hearing lab in the Department of Computer Science at the University of Sheffield. The expected

time for a test is maximum 30 minutes and no breaks should happen during the test so, please be prepared before the experiment starts. However, in case of emergency circumstances or personal reason you can quit the test at any moment.


None of the sounds or scenes inside the computer simulation could be considered harmful for a human. Acoustic noise inside simulation is less than 80db (a vacuum cleaner = 80db and Threshold of discomfort = 120db).

## Procedure:

If you agree to be in this study, you will be asked to do the following:

1. Mobile phones and other devices should be switched off or left outside the sound-attenuating booth.
2. Please avoid moving the chair or table or making any movements that may cause noise during the test
3. You need to wear a headphone and a head mounted microphone which is used for communication and recording purposes. When both participants are ready, the experiments will start.
4. Each volunteer should agree to play either the role of a Firefighter or a TFL.
5. Please avoid of expressing any information that would make it possible to identify you or the other participant during the recordings.
6. At the end of first experiment, volunteers will be asked to change their roles and repeat the experiment with their new roles.

## Firefighting role:

- You need to sit in front of the computer screen and do the practice session.
- You will play with a software in the form of computer game which simulates an incident scene.
- The experiment will starts with pressing the **Connect** button in GUI.
- After a successful connection you will hear a **beep** sound.
- Use your keyboard to move your avatar. You can use the **arrow keys**  on the keyboard to move **forward-backward** and **left-right**. If it is necessary you can use **Space-bar** to **jump** over some obstacles.
- The experiment will automatically be terminated after 30 minutes and again the **beep** sound will be heard at the end.
- Your **goal** is to **explore the incident scene as much as you can** and **accurately report your observations to the TFL**.
- It is expected from you to explain every action that you take (e.g. taking the left door and moving to the next room) and what you can see in which room to the TFL.
- The explanations need to be as clear and accurate as possible to give **clear view of the situation** and **your location** to the TFL. (e.g. what you can see and in which room you think you are)
- Your explanation will be used by the TFL to draw a map of incident area and estimate your location in that.
- You should also reply to the requests from TFL about repeating or clarifying your situation.
- Since the TFL tries to keep track of your movement in all places that you have visited, **you can ask for any potential help (e.g. direction) in case you find yourself lost** (e.g. When you think you are visiting the same location again).
- You can accept suggestions from TFL about directions to unvisited areas.
- You are not racing against time in this experiment; so, please take your time and explain accurately.

### Task Force Leader role:

- You will be asked to sit behind a desk.
- You will be given a rough plan of the environment (2D paper map) and two blank papers for making notes (if it is required).
- The environment map does not contain any information about the name of the areas/rooms or what is inside the rooms.
- Firefighter's starting point and direction is marked with a →
- When the firefighter press the connect button you will hear a **beep** sound and the experiment starts.
- The experiment will automatically be terminated after 30 minutes and again the **beep** sound will be heard at the end.
- Your goal is to listen to the firefighter's report and **imagine the situation and the firefighter's location** based on her/his explanations.
- You should add **any information** that you notice in the firefighter's explanations (conditions, landmarks etc.) in to your paper map and **draw the firefighter's trajectory** based on your estimation of her/his location.
- You can ask for repeating or for clarifying her/his situation if you missed what s/he said, but try not to interrupt her/his work too much.
- You should try to keep track of the firefighter movement.
- You are not racing against time; so, please be patient and let the firefighter carefully explore the environment.
- It is very likely you lose the firefighter exact location. It is not a problem, try to estimate it with what you have. Do not try to ask the firefighter to go back or search the map for you.

### Confidentiality:

Participants' names will be associated with an experiment identification tag, and this tag will be associated in turn with the participants' recorded voice and responses recorded via the keyboard. A file will be maintained by the lead researcher that relates the experiment identification tag to the participants' names in order that participants may be asked to return should any experiments contain anomalous results. Once the experiment is completed, the list of names and participant identification tags will be destroyed.

### Why have I been chosen?

You have been chosen because you are a (self-reported) normal speaking and hearing listener, who is a native speaker of English with standard southern British accent.

### Disadvantages and risks of taking part:

No possible disadvantages or risks are envisaged.

### What if something goes wrong?

In the first instance you should contact the Principal Investigator (contact details are given at the end of this document) should you wish to raise a complaint. However, if you feel your complaint has not been handled to your satisfaction you can contact the University's 'Registrar and Secretary'.



### What will happen to the results of the research project?

The outcome of this study may form part of one or more scientific publications; you will be entitled to copies of any such publications. You will not be identified in any report or publication. The data collected during the course of this study might be used for additional or subsequent research.

### Who is organising and funding the research?

This research is supported financially by the University of Sheffield.

### Who has ethically reviewed the project?

This project has been ethically approved via the University of Sheffield's ethics review procedure.

### Consent:

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep, and will be asked to sign a consent form. You can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason. Information that would make it possible to identify you or any other participant will never be included in any sort of data corpus or report. You can write your email address at the end of the consent form. So you might be invited again for repeating this experiment.

### Questions and contacts:

At this time you may ask any questions you may have regarding this study. If you have questions later, you can contact:

**Supervisor:** Professor Roger Moore: [r.k.moore@sheffield.ac.uk](mailto:r.k.moore@sheffield.ac.uk)

**PhD student:** Saeid Mokaram: [s.mokaram@sheffield.ac.uk](mailto:s.mokaram@sheffield.ac.uk)

**Date:** Date (fixed)

### By signing below, you are agreeing that:

1. You have read and understood the participant information sheet.
2. Questions about your participation in this study have been answered satisfactorily.
3. You are aware of the potential risks (if any).
4. You are taking part in this research study voluntarily.

-----

Date dd / mm / yyyy

Participant's Name and Signature

## C.2 Personal information form

**The University of Sheffield**

### **Personal Information**

#### Participant's Personal Information

**First name:**

**Last name:**

**Age:**

**Sex:**

**Email:**

#### Participant's Accent Information

o you consider yourself as a native English speaker?    Yes ☐ / No ☐

o you believe you have a standard/southern English accent?    Yes ☐ / No ☐

you believe you have any specific accent please mention your accent:

Where did you grow up?

## C.3 Consent form

**The University of Sheffield**

### Consent form

## The Role of Voice Communications in the Search and Rescue Environment

---

### Researchers:

Lead researcher: Saeid Mokaram ([s.mokaram@sheffield.ac.uk](mailto:s.mokaram@sheffield.ac.uk))

Supervisor: Professor Roger Moore ([r.k.moore@sheffield.ac.uk](mailto:r.k.moore@sheffield.ac.uk))

1. I confirm that I have read and understand the information sheet dated 25th September 2014 explaining the above research project and have had the opportunity to ask questions about the project.
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline.
3. I understand that my responses will be kept strictly confidential. I give permission for members of the research team to have access to my anonymised data. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.
4. I agree for the data collected from me to be used in future research.
5. I agree to take part in the above research project.

**Name of participant** \_\_\_\_\_ **Date** dd / mm / yyyy

**Participant's email address:** \_\_\_\_\_

**Signature** \_\_\_\_\_  
(or legal representative)

**Lead researcher:** Saeid Mokaram

**Date** dd / mm / yyyy

**Signature** \_\_\_\_\_  
To be signed and dated in presence of the participant.

**Copies**

Once all parties have signed this, the participant should receive a copy of the signed and dated participant consent form, the letter/pre-written script/information sheet and any other written information sheet provided to the participants. A copy of the signed and dated consent form should be placed within the project's main record (e.g. a site file), which must be kept in a secure location.

**Please address any queries to:**

Saeid Mokaram ([s.mokaram@sheffield.ac.uk](mailto:s.mokaram@sheffield.ac.uk))

Participant ID tag: \_ \_ \_ \_ \_

# Appendix D

## SSAR Corpus Transcription Guidelines

---

This appendix presents the SSAR corpus transcriptions guidelines sheet which is provided for the transcribers. It presents a brief instruction how to use the *Transcriber* (Lieberman et al., 1998) software, how to segment the utterances and transcribe words and noises.



The  
University  
Of  
Sheffield.



# Guidelines for SSAR Speech Transcriptions

Transcription supervisor: Saeid Mokaram

Version 1.4 May 2015

**\*\* Please note that you must read the entire document before starting to transcribe. \*\***

## Background to Project

The Sheffield Search and Rescue (SSAR) corpus is a two-party human/human conversational speech corpus which was made based on an abstract communication model during search process in a simulation environment. In this model the main speaker plays the role of a First-Responder by exploring a simulated environment (Fig 1) and reports his/her observations back to a Task-Leader (second speaker). The design of SSAR corpus targets the automated extraction and understanding of valuable information in human/human conversations in the context of crisis response communication scenario.

In projects related to automated speech recognition and spoken language understanding, availability of high quality annotated databases is a critical issue. Therefore, an effort has been made to design, collect and annotate a two-party human/human conversational speech which includes 12 hours of audio, word level annotation, speaker location and actions information for each conversation. The aim is to make the SSAR corpus widely available for the research community, thereby contributing to the research infrastructure in the field. This work is a part of a PhD research which is started in December 2012 and is expected to finish at this stage by the end of 2015 and may continue afterwards.



Figure 1: Simulation environment

## TranscriberAG software

TranscriberAG is a tool for segmenting, labelling and transcribing speech documents. Here is the link to the TranscriberAG web-page: <http://trans.sourceforge.net/en/presentation.php>

You can find Linux, Windows and Mac OS versions of TranscriberAG in "TranscriberAG" folder or download the latest version from the TranscriberAG web-page. For installation please follow the "binary installations" manuals in this link: <http://trans.sourceforge.net/en/install.php>

After installation you need to load the provided configuration file ("config.cfg") by following "Options->Load\_configuration\_file..." in TranscriberAG menu.

TranscriberAG is a general purpose transcribing software therefore not all its functions are required for transcribing the SSAR corpus. A list of functions that are required is provided in the Appendix of this guideline sheet.

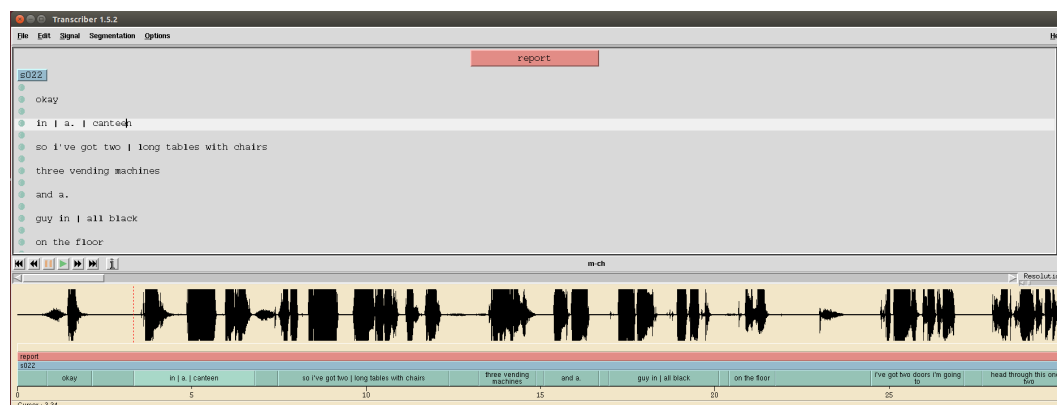


Figure 2: The interface of TranscriberAG. The gray area is where the text transcriptions are inserted. The yellow area shows the wave form of the speech signal. The green horizontal bar shows the speech/non-speech segments and below that is time.

## Ask for help

Anytime you need help, you can contact Saeid Mokaram:

Email: [s.mokaram@sheffield.ac.uk](mailto:s.mokaram@sheffield.ac.uk) ; Mobile: 07453678187

Address: Speech and Hearing research group, room 141, Department of computer science, University of Sheffield, S1 4DP

## Transcription Guidelines

The following flow chart (Fig 3) outlines briefly the basic steps of the transcription process. Each of these steps is then explained in detail below.

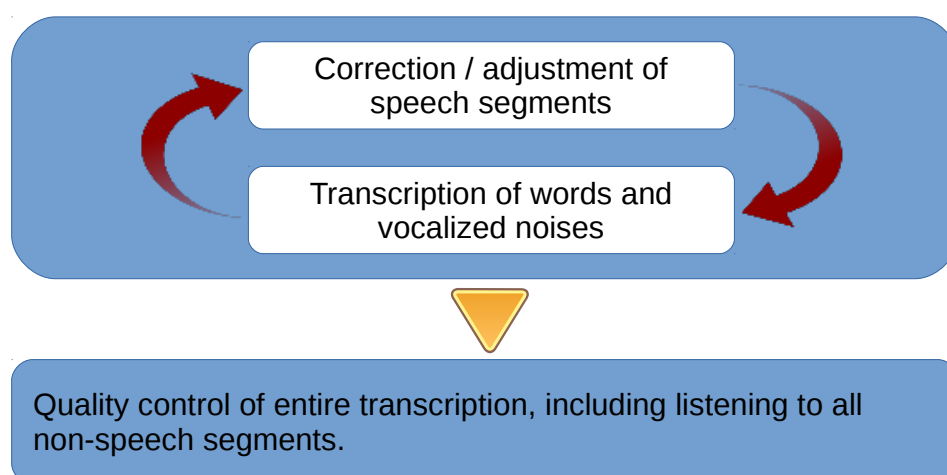


Figure 3: basic steps of the transcription process

### 1 How to Open a Transcription

You may have received either `"m-ch-command_file.txt"` or `"s-ch-command_file.txt"`. It includes lines of commands with the format such as (in Linux version) `"trans map1/start-room6/s022-s021/16kHz_16bit/m-ch.wav map1/start-room6/s022-s021/transcript/m-ch.tris"`

1. First you need to open a command window (terminal Linux/Mac or command prompt Windows).
2. Then change your current directory to the "data" folder (e.g. `"cd ~/trans/data"`).
3. What you need to do is just copy one line at a time and past it in the Terminal. After pressing Enter the TranscriberAG starts and automatically loads an audio file and its corresponding automatic segmentation and transcription file.
4. Now one audio document is ready for transcription.
5. Remember to save (**Ctrl+S**) frequently during transcription.
6. After finishing one transcription go back to step 3 until finishing all the lines of commands.

It is recommended that you make a copy of `"m-ch-command_file.txt"` and `"s-ch-command_file.txt"` files so when you finished with one audio document you can remove the corresponding line of command in the copied file. It will help you keep track of the audio documents which are not transcribed yet.



All the audio documents are auto-segmented and auto-transcribed. This is just to make the task slightly easier for you; however it doesn't mean that they are correct and reliable. So you need to check the transcriptions and correct them according to the presented standards in this guideline.

This is a two-party conversation similar to a telephone talk. The main speaker voice and the second speakers voice are recorded on m-ch and s-ch channels respectively. This means that you will hear one speaker at a time while you are transcribing. These two types of audio documents are slightly different. The difference is that all the “m-ch”s are mainly speech, while all the “s-ch”s are mainly silent with just very short speech parts (such as confirming, asking questions etc.). Transcription supervisor will allocate you with either m-ch or s-ch type of data.

## 2 Segmentation

Each audio document must be broken into small segments, ideally separating speech from non-speech/silence regions. This task can be very time consuming, and for this reason we have used an automatic pre-segmentation system to do a first pass on the data.

As you listen to the audio documents the main task here is to check and adjust the boundaries of the segments. The segment boundaries should be adjusted to ensure they accurately indicate when the speech starts, when it stops (ensuring not to cut any sounds off at the beginning or end of the segment).

Automatically detected segments may need to be broken into smaller segments, or have their boundaries adjusted. Silence/pause is really the main determiner, perhaps followed by syntax.

These smaller segments are often called utterances. The speech can be broken into chunks based on pauses within the speech stream, preferably at syntactic boundaries corresponding to sentences.

At first this may seem a little difficult as clear grammatical sentences are not often present in natural speech. However, once you become more experienced doing speech transcriptions the task of where to place segment boundaries becomes clearer and quite natural.

The general rule is that if you have to think too long and hard about where to place a segment boundary, you probably don't need to put one in.

### Important things to remember about segmentation:

1. Each segment should be padded by a small buffer of silence ( $\frac{1}{4}$  -  $\frac{1}{2}$  second) on both sides (before and after the speech) if possible.
2. Breakpoints should be inserted at natural linguistic points in the utterance such as sentence or phrase boundaries to the extent possible.
3. Generally most of the speech segments (utterances) are only a few seconds and not more than 1 minute in length.
4. Utterances may start or end with Vocal-Noises such as “aspiration”, “cough”, “laugh” and etc.; if there is absolutely no silence between them and speech region you need to include them in speech segment (the vocal-noise is mixed with part of speech). You can ignore those vocal noises that are not loud and you can hardly hear them.

5. Vocal noises can be at the middle of the utterances with no or a small silence gap before or after them. You don't need to break the utterances because of them. You just need to add proper Vocal-Noise-Tags in their place.
6. If a segment only contains vocal-noise it doesn't need to be annotated. Leave it as blank similar to a silence segment. If it is already automatically annotated with some words or tags, you need to remove them.

### 3 Words and Vocal Noise Tags

In this section we explain the steps and guidelines for transcribing the speech within a given time segment. In general, we break noises that are made using the mouth into two categories, Words and Vocal-Noise-Tags.

Most of the speech you encounter can be transcribed into words with standard orthographic representations, such as you would find in a dictionary. But you will also encounter several other types of “verbal” events which will also need to be transcribed. These include words that may not appear in a standard dictionary but that are common in speech, such as reduced forms like “dunno” or “wanna”, or acknowledgements like “uh-huh”, or pause-fillers like **um, em, erm, uh, eh, yep, yep, yeah, yup...** . In addition, when speech is broken off, there may be word fragments. Finally, there will be vocal sounds, like laughs and coughs and sighs which do not have usual lexical representations. For this transcription effort, we consider all except this last category to be “words” and transcribe them as such, using a standardized set of spellings; the members of the last group are instead transcribed using special tags more fully described below.

It is important to remember that (where possible) the transcription should be an accurate record of what was actually said, and speakers should not be corrected to make their speech more grammatically correct. For example if the participant says “I dunno” this should be transcribed as it is heard, not as “I don’t know”.

In case you encounter spoken phrases which you are not confident what written form to use, write what you think is correct in parentheses ( e.g. (whatcha) ) so later they will be checked by transcription supervisor.

Vocal-Noise-Tags describe sounds that are made using the mouth (or nose) but that do not have standard lexical representations. In these transcriptions, we have reduced the number of these which will be annotated to four, each of which has a simple symbolic representation. These will be:

Vocal Noise Tags	Notes
@	aspiration
%	cough, throat clearing
\$	laugh
#	other prominent vocal noise (e.g. creaky voice, yawn, tongue click, etc)

Table 1: Vocal-Noise-Tags

The following text provides an example:

*okay % i'm in the first room and it's a kitchen*

*\* It will probably be useful to make a list of these symbols and keep them handy when you are transcribing, perhaps on a sticky note glued to your monitor.*

### Important information about transcribing words and vocal noises

1. Transcribe verbatim, without correcting grammatical errors, e.g. "I seen this room before".
2. Standard spoken language should be transcribed as it is spoken, e.g. "gonna" not "going to", "wanna" not "want to", "kinda" not "kind of", etc.
3. Avoid word abbreviations, i.e. "doctor" not "Dr", and "mountain" not "Mt".
4. Remember to watch for common spelling confusions like "its" and "it's", "they're" and "there" and "their", "by" and "bye", "of" and "off", "to" and "too", etc. which are common in the automatic transcription.
5. Mispronunciations: if a speaker mispronounces a word and you know what word was intended, transcribe the word as it should be spelled and mark it with an asterisk after the last letter, e.g. spaghetti\*. If you do not know what word was intended, transcribe what you hear and mark it with parentheses, e.g. (fligop).
6. Spell out number sequences, e.g. "forty four" not "44".
7. Acronyms should be spelt as they are pronounced, e.g. "nasa", "t\_v\_" or "d\_v\_d\_" (be careful it is automatically transcribed like "t. v." ; Please change it to t\_v\_ ).
8. If a speaker does not finish a word, and you think you know what the word was, you can spell out as much of the word as was pronounced inserting a single dash as the last letter of the word, e.g. "I'm in a kitch-".
9. Punctuation should be limited in the corpus. You should only use full stops and question marks to punctuate a 'sentence'. Most of the speech segments are utterances (not sentences) therefore, most of them do not need full stops. Some of the utterances are the final segment of a 'sentence'. If you could detect them use full stops or question marks.
10. Some times there are small silence gaps at the middle of utterances (about ¼ to ½ second). Some of them are automatically detected and marked with | however, if there is an undetected silence gap you need to add | at its place or if there is a miss-detection you need to remove the | or replace it with proper words/vocal-noise-tags. In the example below, the speaker made a small pause after saying "there are" and then immediately starts the "two fridge freezers"; so a | indicates this small silence.

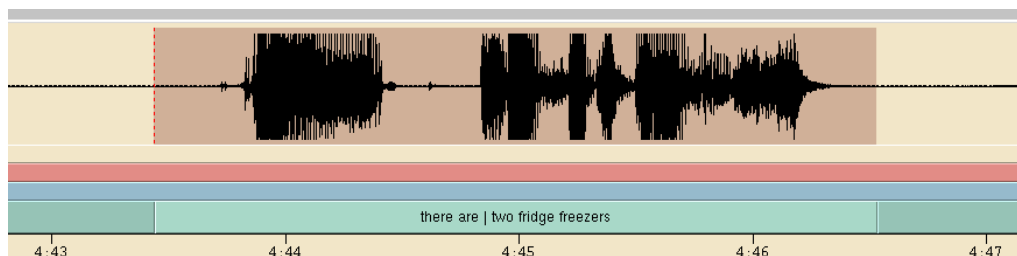


Figure 4: “there are | two fridge freezers”

7. Where the speaker interrupts himself to correct or restart or repeat, you need to repeat it as it is. For example in “two one one large cupboard area” which the speaker repeated “one” to correct himself.
8. In similar cases when there is a small silence gap use a | . For example: “I just meant | I mean ...”.
9. Please note that we are using British standard spelling throughout the transcriptions, i.e. colour vs. color, realise vs. realize.

## Practical Procedure

This section outlines the practical procedure for annotating the speech in transcriptions.

You should have just adjusted the time boundaries for a given segment. By the time that you have done this, it is likely that you have understood a lot of what is said in that segment. For this reason, it is a good idea to transcribe this segment before moving on to the next one. The following steps will help you to accurately transcribe the speech.

***Remember we are interested in a balance between quality, consistency and accuracy.***

**The procedure below explains how to transcribe each speech segment:**

- 1) Is what you heard a Vocal-Noise or silence?

Yes: Type proper Vocal-Noise-Tag or | for silence.

No: Go to Question 2 .

- 2) Did you understand what was said?

Yes: Go to Question 3.

No: Listen again, if you think you might know what the word is but are very uncertain, type that word inside parentheses, e.g. (egg). If after the second listen you still don't know what a word is, mark the transcription with a '(??)', i.e. using question marks inside the parentheses, where that word should go. Move on to the next word and start again at Question 1. Parentheses should not be used for any purpose other than to indicate uncertainty in these transcriptions. You should in fact mark anything you

are uncertain about with parentheses, so that transcription supervisor will easily be able to identify any problem areas.

3) Are you sure about the its spelling (British standard spelling)?

Yes: Go to Question 4.

No: Type the word as you think is correct inside parentheses, e.g. (apple).

4) Does it seem to be a word fragment or unfinished word?

Yes: Transcribe as many of the phonemes as you hear, marking that the word has been cut off with a '-' as the last letter. If you are uncertain about the phonemes you hear, mark the whole thing with parentheses. e.g. " I'm in a kitch- ".

No: Go to Question 5.

5) Is that an acronym:

Yes: Type as they are pronounced, e.g. "nasa", "t\_v\_ " or "d\_v\_d\_" (be careful it is automatically transcribed like "t. v." ; Please change it to t\_v\_ ).

No: Type the word and move on to the next word and start again at Question 1.

## Appendix

The following table provides details of the TranscriberAG functions that are required when you are transcribing the SSAR corpus.

Functions	Notes
Segmentation	You will note that the transcription has already been divided into speech/non-speech segments for each channel.
Save: <b>Ctrl+S</b>	Remember to save frequently during transcription.
Select a section of the transcription: <b>Leftclick on green bar</b>	If you left click your mouse on one of the sections in the green transcription channels, that section of the audio channel will be highlighted. If you play the audio now it will only play the audio for that section.
To select several segments together: <b>Shift+leftclick on green bar</b>	If you left click your mouse on one of the sections in the green transcription channels, that section of the audio channel will be highlighted. If you play the audio now it will only play the audio for that section.
Change a segment boundary: <b>Ctrl+leftclick+drag boundary</b>	If you find that the automatic segmentation boundary needs adjusting (perhaps it cuts off the end of the last word that the person says) then you should hold down CONTROL and LEFTCLICK and DRAG the boundary to its new position.
To add a section boundary: <b>Enter</b>	If the automatic segmentation has missed a segment (for example a section of silence) you should first CLICK on the green channel, then position the cursor at the point in the audio where the section break should appear, and then hit ENTER to add a break.
To delete a section boundary: <b>Shift+Backspace</b>	If the automatic segmentation has incorrectly split what should be a single segment into two segments, you should remove the border between them. LEFTCLICK in the second segment (on the green bar) and hit SHIFT and BACKSPACE.
Play/ Pause: <b>Tab</b>	The TAB key will toggle between play and pause. However, if you have a single segment selected, only that segment will play when you hit TAB. If you then hit TAB again (after it has played that segment), it will play that segment again. (It may be worth noting that you can modify playback options using the appropriate pull-down menu. It is sometimes helpful to switch to continuous playback or pause-and-continue mode, rather than stopping at segment boundaries.)
Quit/ Exit: <b>Ctrl+q</b>	Quit/ Exit
<b>Up/Down cursor keys</b>	Move between segments

Table 2

Annotators may choose to print this summary table and keep it nearby (perhaps attaching it to their computer screen) while they become more familiar with the software.

Task	Short-cut
Save	<b>Ctrl+S</b>
Select section	<b>Leftclick on green bar</b>
Select several segments	<b>Shift+leftclick on green bars</b>
Change segment boundary	<b>Ctrl+leftclick+drag</b>
Add section boundary	<b>Enter</b>
Delete a section boundary	<b>Shift+Backspace</b>
Play/ Pause	<b>Tab</b>
Quit/ Exit	<b>Ctrl-q</b>

Table 3

Example	Description
(egg)	You are not sure it is actually pronounced “egg”
(??)	You don't know what is pronounced.
	There is a short silence gap inside an utterance.
t_v_	Speaker pronounced TV
nasa	Speaker pronounced NASA
kitch-	uncertain about the rest of phonemes you hear OR it's a word fragment (“kitchen”)

Table 4





# Appendix E

## Examples of Manually Estimated Maps

---

This appendix presents examples of all four maps which are estimated by a participant (Participant-ID s004) in the role of a task leader. In total, the SSAR includes 96 hand drawn topological estimations, one for each recording session (conversation).

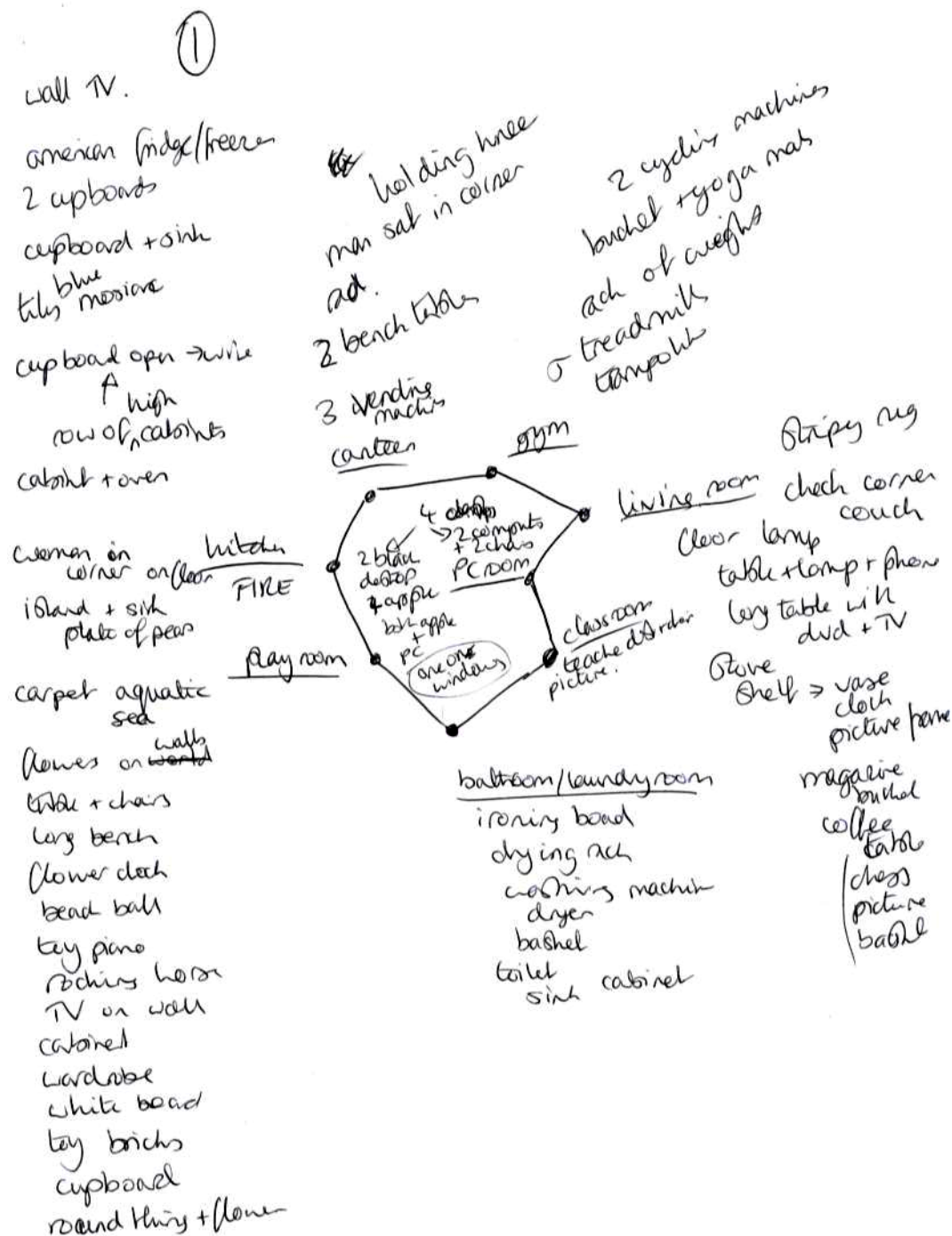


Figure E.1 A typical example of hand drawn topological map of the Map<sub>1</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004).

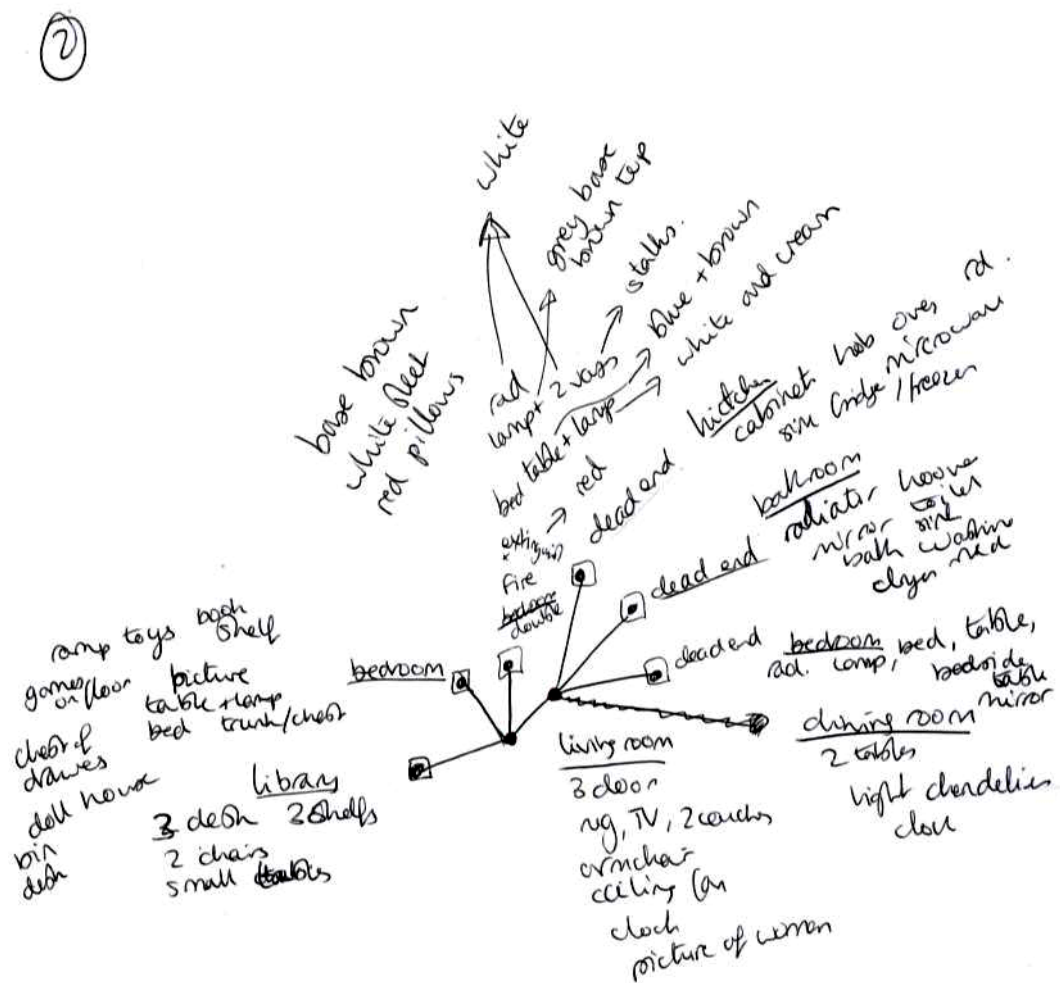


Figure E.2 A typical example of hand drawn topological map of the Map<sub>2</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004).

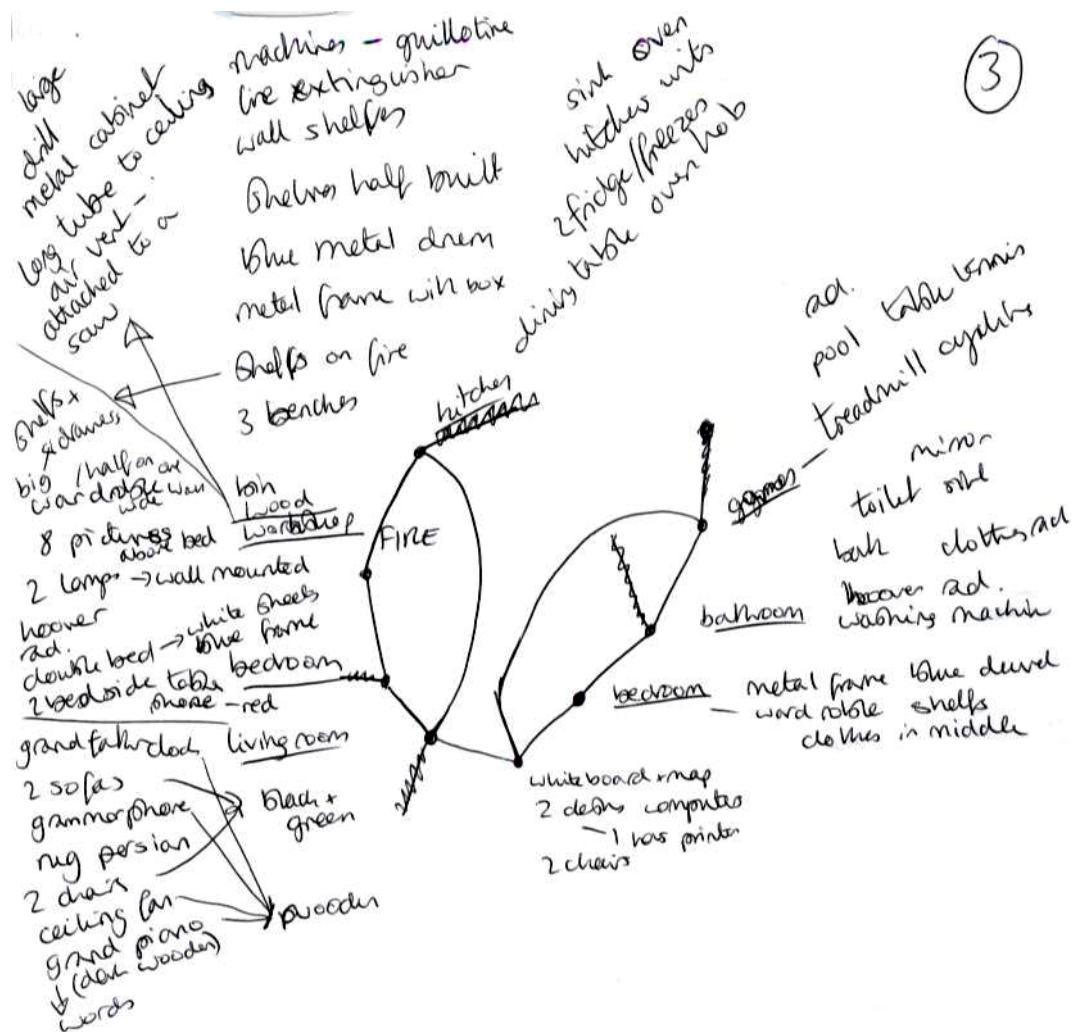


Figure E.3 A typical example of hand drawn topological map of the Map<sub>3</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004).

④

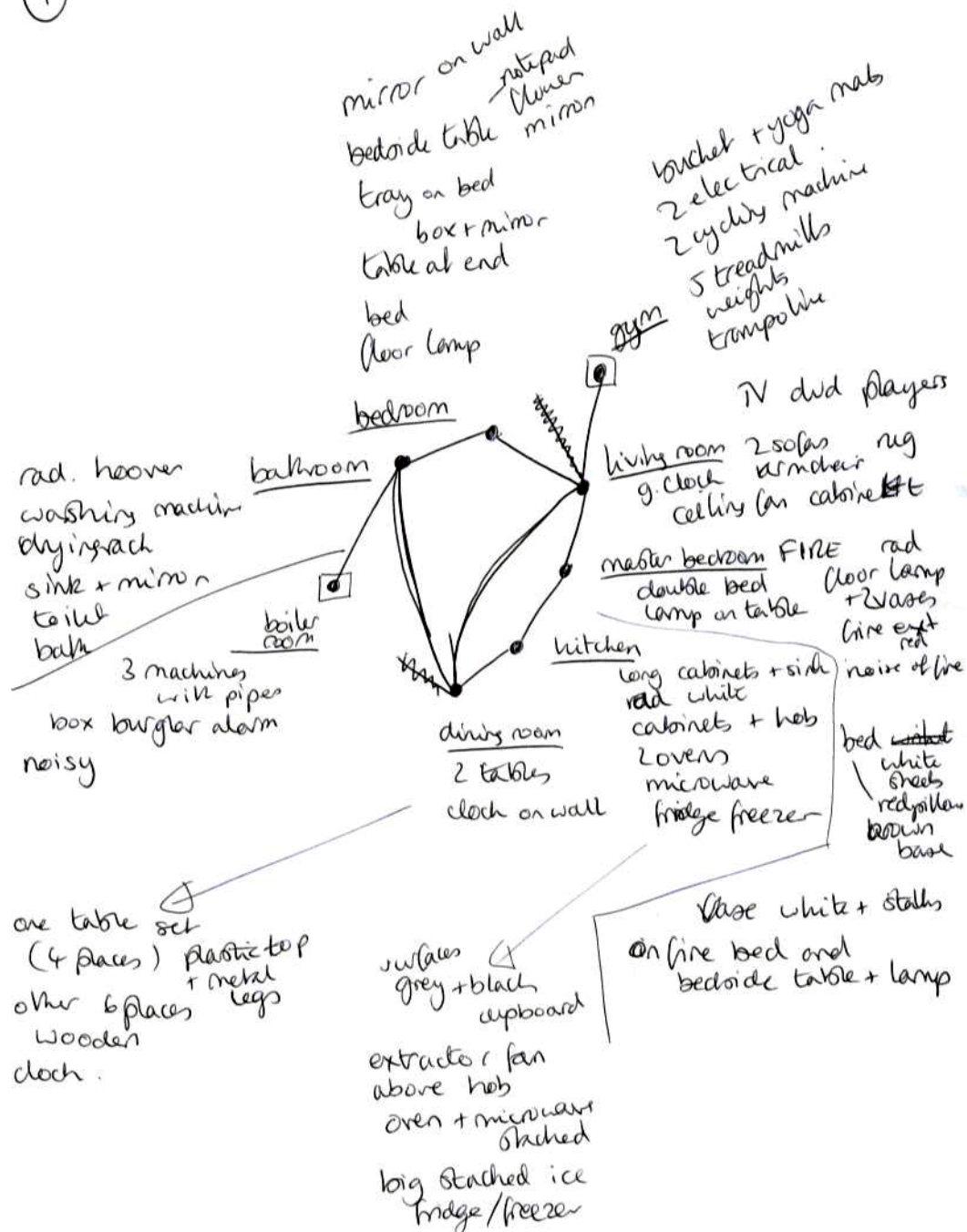


Figure E.4 A typical example of hand drawn topological map of the Map<sub>4</sub>. This example map was estimated by a participant in the role of a task leader (Participant-ID s004).



## ASR Performance at Different SNRs and LMSFs

---

Figure F.1 illustrates the baseline ASR (cf. Chapter 5) WER landscape at different SNRs and all LMSFs. Figure F.2 shows how the LMSF was shifted towards more weight on the language model in high acoustic noise. This figure presents the ASR performance normalized at each SNR between zero and one.

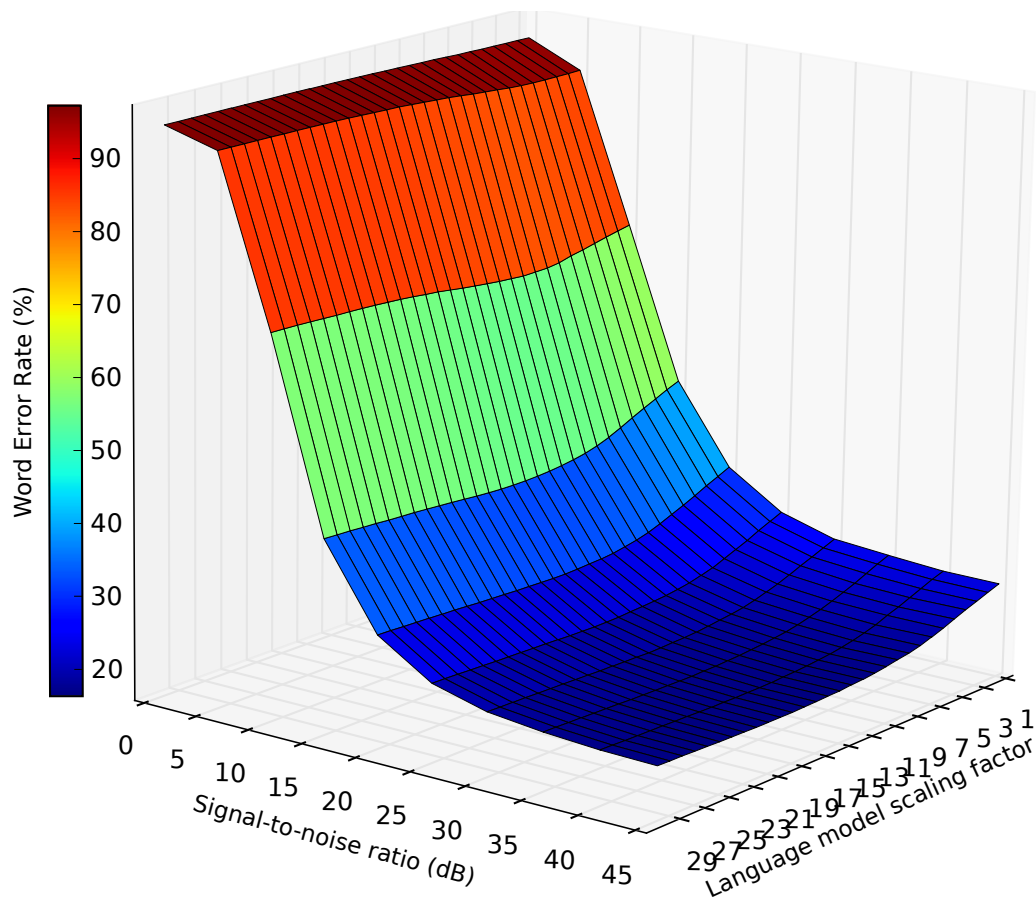


Figure F.1 *The baseline ASR WER landscape at different SNRs and all LMSFs*



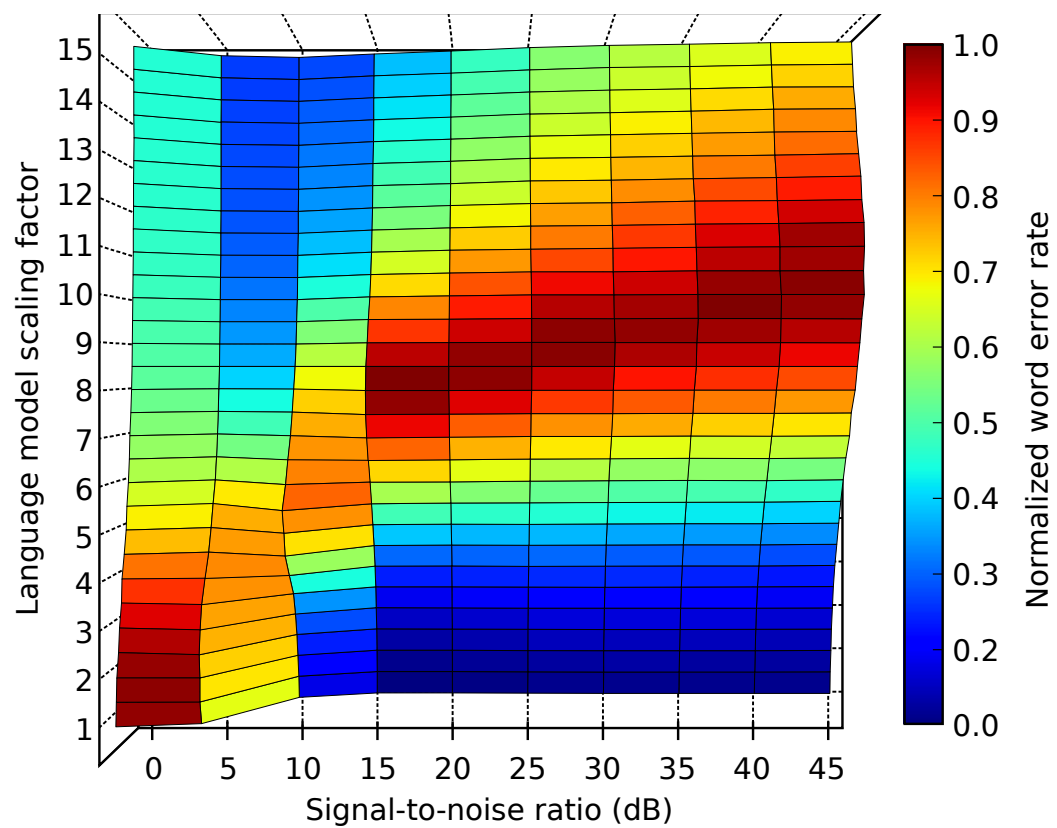


Figure F.2 *The baseline ASR performance landscapes at different SNRs and all LMSFs. The ASR performance is normalized between zero and one at each SNR.*



## List of Publications and Presentations

---

Some parts of the work presented in this thesis have been published in conference proceedings as given below:

1. Saeid Mokaram and Roger K. Moore, “Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario”, in *Interspeech*, 2015. (Oral presentation)
2. Saeid Mokaram and Roger K. Moore, “Speech-Based Topological Map Estimation in a Simulated Search and Rescue Environment”, in *NIPS 2015, Workshop on Machine Learning for Spoken Language Understanding and Interaction*, 2015. (Poster presentation)
3. Saeid Mokaram and Roger K. Moore, “The Sheffield Search and Rescue Corpus”, in the IEEE International Conference on *Acoustics, Speech, and Signal Processing* (ICASSP), 2017. (Poster presentation)
4. Saeid Mokaram, Hamideh Kerdegari, Christina Georgiou, Roger K. Moore, Tony J. Prescott, Tony J. Dodd, “Search and Rescue 2020”, University of Sheffield Engineering Symposium 2013 group project. (First prize in group poster competition).
5. Saeid Mokaram and Roger K. Moore, “High-Level Context for Improving Automatic Recognition of Conversational Speech”, Submitted to *Interspeech*, 2017.

