Speech-Based Topological Map Estimation in a Simulated Search and Rescue Environment

Saeid Mokaram
SpandH, Department of Computer Science
The University of Sheffield
Sheffield, S1 4DP, United Kingdom
s.mokaram@sheffield.ac.uk

Roger K. Moore
SpandH, Department of Computer Science
The University of Sheffield
Sheffield, S1 4DP, United Kingdom
r.k.moore@sheffield.ac.uk

Abstract

In a Search & Rescue scenario, a first responder's (FR) spoken description can be viewed as a verbal annotation of the incident scene. This paper presents a method for interpreting such descriptions as a topological representation of the incident scene. However, in contrast to a traditional approach using units of meaning, our approach uses a topic-based perspective since it offers the potential of being robust to high error rates in the automatic recognition of noisy speech. We thus frame the landmark detection problem in topological mapping as a topic segmentation task. New nodes are introduced to the map by identifying the changes in the content of spoken reports as an indication that the speaker has moved from one location to another, and text vectorization techniques are used to compute the correspondence between pairs of nodes and to estimate the topological map. A goal-oriented human/human conversational corpus was collected involving spoken communication between a FR and a task leader in a simulated search environment. Experiments on manual and automatic transcriptions with different levels of word error rate confirm the low sensitivity of this method to highly imperfect speech recognition output.

1 Introduction

Whilst spoken language understanding (SLU) 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 a clear need for applications such as Search and Rescue (SAR) operations. Speech is the single most important source of situational information during crisis response. It is widely used for transferring critical information about the incident scene between First Responders (FR) and Task Leaders (TL) [1]. Automatic extraction and integration of these information into the central information management system has the potential for reducing the risk of human related errors in large and fast moving SAR operations. The importance of such automatic estimations based on speech communications has been envisaged in the observational-speech-system [2]. Yet, technical difficulties such as high word error rate (WER) in automatic speech recognition (ASR) transcripts and understanding spontaneous human/human communications present major challenges for implementing such a system.

Reliable information about the lay of the land is known to be one of the main enhancing factors for situation awareness within the SAR context [3]. In general the internal representation of a physical environment can be classified as metric-based or Topological-based maps (T-map) [4]. While a metric map represents the geometric entities of an environment as exact locations, a T-map represents the structure of a physical environment as an abstract graphical model consisting of nodes and edges [4, 5]. Estimation of a metric-map from spoken reports requires highly sophisticated ASR and SLU systems for interpreting detailed information about the location of environmental entities. Prior to this challenge, it is also very unlikely to find detailed metric information in FRs' explanations which are generally about main locations and events during the explore (search) phase. The topological representation methodology is similar to the environmental perception and interpretation of human beings [6] which makes it more applicable in the speech-based mapping problem. Followed by our previous work in estimating FRs' location based on their spoken reports [7], this paper presents a

speech-based T-map estimation approach based on the automatic transcripts of speech communications in a simulated SAR scenario.

The explanations of FRs about their observations and actions are highly associated with their location. We have shown that [7], topic segmentation techniques can provide an estimation about when the FR has moved from one room into another. Looking from a topic segmentation perspective and utilizing the *watershed+Vectorization* segmentation technique for a 1D signal [8] provided a significant amount of robustness to the high WER in ASR transcripts of spontaneous human/human spoken language communications. Here we thus frame the landmark detection problem in T-map building as a topic segmentation task. New vertices are introduced to the graph of the T-map by identifying the changes in the content of FRs' spoken reports as an indication to that the speaker has moved from one particular location to another. The correspondence between pairs of nodes is estimated by first describing their entire segment of utterances in the vector space model (VSM) and then computing the *cosine* similarity of their vectors as a measure of their correspondence. The entire segment is comparing in VSM in order to reduce the perceptual aliasing posed by the ASR errors and improve the distinctiveness of nodes.

Experiments on the manual and automatic transcriptions with different levels of WER (32.4% & 41.6%) confirm the low sensitivity of this method to highly imperfect speech recognition output. The results show the capability of this system in extracting on average 52.52% of the total information content of the spoken report about the structure of the environment on manual transcriptions and on average 45.15% and 43.46% on clean and noisy speech data respectively.

2 Topological environmental modelling

T-maps have been mainly studied by cognitive theories of space [6] and mobile-robot mapping [4]. The standard definition of a T-map [5], describes it as a graph which its vertices or nodes represent certain distinguishable places in the environment (landmarks) and the edges or links between them indicate the connections between their corresponding locations. In contrast to metric-maps, these light-weight maps can represent higher-level of semantic knowledge such as objects and semantic labelling about the environment.

Automatic T-map building is a well explored area mainly in the field of mobile-robot mapping [4]. As the robot agent explores, it perceives the environment through its sensors. One of the main issues in T-mapping is to detect when a new node should be added in the map. Some of the existing approaches place nodes periodically in either space (displacement) or in time intervals. In some other strategies, 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 landmark detection produces a more compact topology which the nodes can represent higher level of semantic knowledge. Numerous types of range-finder sensors and vision-based methods have been employed in the literature to interpret the environment and identify these topological landmarks [4].

In this process, a sequence of nodes is generated as the agent explores the environment. At this stage, building a T-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 which is also known as the *correspondence problem* in T-mapping. Solving the correspondence problem is made difficult due to *perceptual aliasing* in the environment in which different places may have similar appearance or they may look similar to the system. In another situation, due to *perceptual variability*, a single place visited two times can appears distinct to the system. This may occur because of the viewpoint or illumination effects among other causes. Failure to assess the correspondence between landmarks accurately, increases the ambiguity of the T-map [9]. Current mapping algorithms mostly solve correspondence problem by matching low-level sensor-specific features. Although several approaches [10] (among others) 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 probabilities of loop closing [11]. It is frequently used in vision-based loop closing [12, 13, 14] which the perceptual aliasing and variability are generally higher.

3 Topic segmentation

Topic segmentation is an essential step in understanding and information retrieval tasks. It has been approached in many different ways and most of them are sharing the use two basic insights either individually or in combination. The first is that, a change in topic will be associated with the introduction of a new vocabulary [15]. 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 or prosodic cues [16, 17, 18]. 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.

Different approaches have been introduced both for content-based and boundary-based. The TEXT-TILING system [19] proposed to use a computation of similarity. It is inspired from the classical approaches in the information retrieval domain such as TF-IDF. In TEXT-TILING system the content of a sliding window is compared before and after each possible boundary. Significant local minima in the lexical cohesion were considered as an indication for topic boundaries. The segmenting task in the SEGMENTER [20] is defined as finding the boundaries on a representation of text as weighted lexical chains. Utiyama and Isahara [21] applied a HMM based statistical approach to measure lexical cohesion with the help of language modelling. DotPlotting [22] used clustering on the similarity matrix between candidate segments. To decide if the topic has changed or not, these approaches rely on word repetition for computing some kind of similarity.

Segmentation task on different genres of speech can be more challenging depending on the structure of the discussion. A human-human spontaneous dialogue is generally much less well-structured and topics can be revisited or interleaved. ASR WER is also significantly higher on spontaneous speech, and all these make this segmentation task more difficult in comparison to more constrained genres such as monologue [23]. In order to deal with short segments with very few common words, Guinaudeau et al [24] integrated the semantically related terms to the HMM segmentation model to extend the description of the possible segments. Claveau and Lefevre [8] introduced the *Vectorization* technique which makes it possible to match text segments that do not share common words. This is especially useful when dealing with high WER in ASR transcripts. They also adopted the *watershed* transform which is a famous morphological method for image segmentation [25] and achieved a high quality of segmentation on transcripts of TV broadcasts.

4 Experimental data

A goal-oriented two-party human/human conversational speech corpus (SSAR corpus) was made based on an abstract communication model between FR and TL during search process in a simulation environment. In this model, FRs goal is to explore the environment and report their observations back to the TL. The recording set-up is visualized in Figure 1. In this arrangement, FR and TL were located in separate quiet rooms. TL could hear FR's reports and in the same fashion, he was also able to talk back for asking or confirming the required information. Given pen and paper and just relying on the FR explanations, the TL was asked to make an estimation about structure of the simulation environment by drawing circles to represent rooms and lines between them to show who they may connected to each other. Inspired from simulation training systems which are being used by some fire departments, a simulated indoor environment was designed.

Figure 1: *Left: The recording scenario, Right: The recording set-up in two separate quiet rooms.*

Four different simulated environment settings were designed in order to have multiple levels of complexity and difficulty. The topological structure of these four map settings are shown in Figure 2 (left). The top-view of the Map_4 settings as an example is also presented on the right. Each map setting consists of 8 rooms. While all the rooms have an identical square shape, different objects and arrangements inside them gives a unique identity to each. Some maps have multiple rooms with the same type; for example Map_2 has two different bedrooms. In total 13 different types of indoor locations ($RoomTypes$) such as *kitchen, bedroom or computer-lab* were simulated in all four map settings. Different types of ambient noises (fire noise, home appliance noise e. g. washing machine) were also simulated which the FR (and also TL) could hear by approaching to the source. Recordings were performed in two separate quiet rooms for avoiding external acoustic disturbances. The two speakers' voice and the environment noise were recorded on separate channels. For annotation purposes, other information about locations and actions of the FR inside the simulated environment was also logged in a computer readable text file.

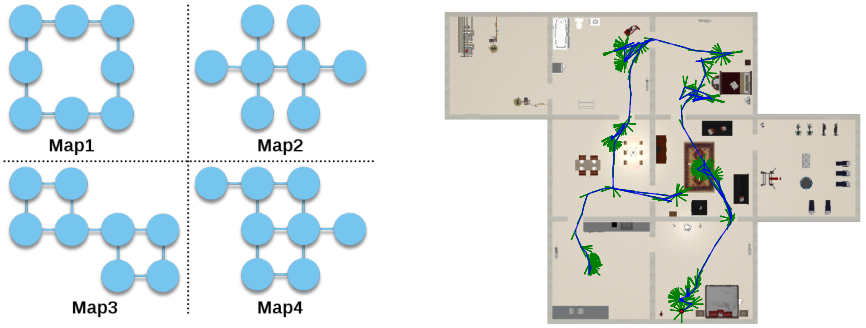

Figure 2: *Left: The topological structure of four different map settings which were explored by each participant. Right: Top view of the Map_4 with the motion trajectory of a participant.*

In total 24 native speakers of British English with southern accent (66.6% Male) participated as paid volunteers recruited through the Sheffield-student-volunteers system. Each participant explored all four map settings which means 96 individual recordings were performed. Majority of the participants in the role of FR explored all the rooms in each map. Although there were about 12.5% who couldn't managed to visit all the rooms in the limited time, in all experiments, the topological structure of the visited parts of the environment was correctly estimated by the participants in the role of TL. This confirms that, the amount of exchanged information through voice channel is sufficient for a human subject to estimate the structure the visited parts of the environment. Therefore, the topological structure related to the visited parts in each recording-set is considered as the ground truth map. Each recording has an average length of ≈ 7.25 minutes. The corpus contains 12 hours of conversational speech with word level manual annotations.

5 Speech-based topological map estimation

The speech-based approach for T-map estimation is designed based on the main principle of T-map making by first, detecting when a new node should be added in the map and then, estimating the correspondence of the recently added node to the previous ones.

5.1 New node detection

Here, the landmark detection approach is used for introducing new nodes into the map. We have shown that [7], looking from a topic segmentation perspective it is possible to segment FR's spoken reports in a way that each segment represent a particular location. It has also been shown that these segments of the report provide enough information for estimating the FRs' locations. Thus they have all the characteristics required to be considered as landmarks. In this method, using the actual transition times from the location information of the FR in the simulated environment, a transition-pivot-document (TPD) was built from all transition-related utterances in the training dataset. This TPD was used as a reference, and a fixed size sliding window (w=5) over the sequence of utterances was compared against it. Using the *Vectorization* ($\vec{V}$) principle [8], both window ($u_{i-k:i+k}$) and TPD were projected into the VSM and the *cosine* similarity between their vectors were computed as follow:

$$D(i) = cosine\Big(\vec{V}(TPD),\ \vec{V}(u_{i-k:i+k})\Big) \quad \forall i \in [1:N],\ \ k = \left|\frac{w}{2}\right| \tag{1}$$

An example of this similarity between the utterances of a conversation and the TPD is visualized in Figure 3. Based on the explained watershed-based segmentation approach in [8], estimations

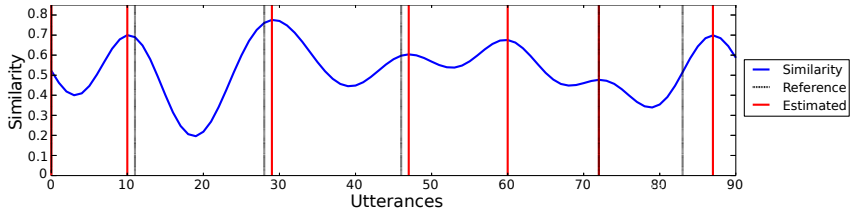

Figure 3: *The cosine similarity distance between the utterances of a conversation and the TPD.*

about the transition times were calculated. These estimations about the transition times divide a long sequence of utterances into smaller segments which each segment can represent a particular

location. By allocating a node for each segment of the utterances, a sequence of nodes is gradually formed. The utterances related to each node are then retained as the fingerprint of the location in which the node is representing.

5.2 Correspondence matrix estimation

At this stage, building a T-map is reduced to determining whether each node in the extracted sequence of locations is a new one or one that has been visited by the FR previously. Here the utterances in each segment are the main source of information for estimating the correspondence between nodes. In this estimation, two situations can be envisaged in which the topological ambiguity (perceptual aliasing and perceptual variability) may occur. First, when a spoken report itself is ambiguous. In other words, the reports themselves are not accurate enough for distinguishing two nodes correspondence. In this situation it should also look ambiguous to the TL. However, since the structure of the environments are correctly estimated by the listeners (TLs) in all the recordings, the information content of each report is enough for a human subject to find its T-map correctly. In the other situation, these explanations can appear ambiguous to the mapping system due to the ASR errors or short segments in the segmentation. In order to reduce the perceptual aliasing posed by the ASR errors and improve the distinctiveness of nodes, each segment of utterances is considered as a whole and then using the *Vectorisation* principle, its bag of N-gram is projected into the VSM. A larger segment of utterances contains more information with higher redundancy and the *Vectorization* technique makes this possible to match text segments that do not share common words or contain errors. More formally, the correspondence (C) between the most recent node (x_n) with all the previously detected ones (x_j) is estimated as:

$$C_{n,j} = cosine\Big(\vec{V}(x_n), \vec{V}(x_j) \Big) \quad \forall j \in [1 : n-1] \tag{2}$$

Figure 4 represents an example of the estimated correspondence matrix ($C^{N \times N}$) (left) for Map_4 and its ground truth (right). An over segmentation can be seen on its ground truth matrix (nodes 10

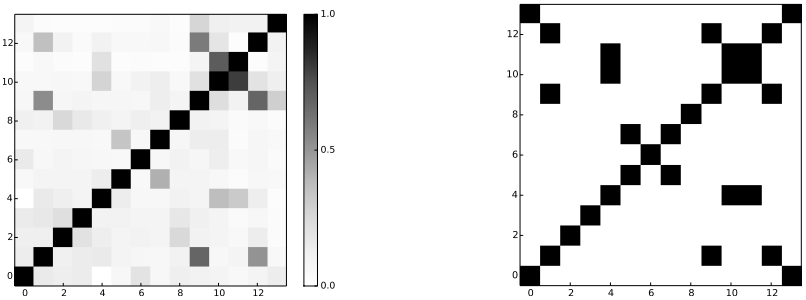

Figure 4: *Left: Example of the estimated correspondence matrix for Map_4, Right: its ground truth.*

and 11 are representing one location). This generates short segments which may contain insufficient data to be correctly compared with the other nodes.

5.3 Experiments

Three sets of experiments were conducted on manual transcripts and ASR transcriptions of clean speech data and speech with the background environment noise. The ASR system used for the experiments was accessed through webASR [26]. The specific system used was a 2-pass DNN-GMM-HMM tandem system trained on 95 hours of speech from 327 British broadcasts. The language model used was a 3-gram based on the interpolation of multiple language models trained on meeting, broadcast and telephone transcripts, with a vocabulary of over 62,000 words. After an initial pass with the speaker independent models a global CMLLR transformation was estimated for each input file and used as a parent transform in the estimation of speaker-based MLLR transformations; the joint CMLLR-MLLR transformations were then used in a final speaker dependent decoding. In the task of transcribing 15 hours of multi-genre television broadcasts [27], this system achieved a WER of 37.5%. In this experiment, this system achieved 32.4% and 41.6% WER on the clean and noisy data respectively.

In each experiment, the K-Fold cross-validation (k=10) was used in order to divide the data into train-dataset and test-dataset. Document vectors were produced by applying the Vectorisation scheme and the introduced node detection and correspondence matrix estimation were applied on the transcriptions of the speech data. The effect of errors posed by the segmentation (node detection) on the correspondence estimation was measured by comparing the overall performance of the system on the auto-segmented transcripts with its performance on the pre-segmented transcriptions which were provided by the actual room transition times.

6 Results and discussion

The performance (P) of the correspondence estimation is estimated as follow:

$$P = \frac{d(R, GT) - d(C, GT)}{d(R, GT)} \times 100 \tag{3}$$

Here the Euclidean distance (d) between the estimated correspondence matrix ($C^{N\times N}$) and its ground truth ($GT^{N\times N}$) is compared against the distance of a same size random matrix ($R^{N\times N}$). This shows the fraction of obtained information from the spoken report out of its total information content about the structure of the environment. Table 1 shows the system results on manual transcripts and ASR output of the clean and noisy speech data. Depending on the topology of each map-setting, the level of ambiguity in the correspondence estimation can vary. Thus, the results for each map-setting are presented independently. For instance, in general, the estimations on Map_1 were more accurate and more information were gained in comparison to other map settings (e.g. Map_2). This is because of its simple circular structure which can be explored (and explained) with much less revisiting the rooms. It is notable in the results that, in spite of a considerable increase in the WER, the system did not receive a high negative impact from that and it is able to extract topological information even in such inaccurate automatic transcriptions. It is important to note that, errors in the segmentation (node detection) can indirectly affect the correspondence estimation by producing short segments. Table 2 shows the effect of such error by comparing the overall performance of the system on the auto-segmented transcripts (Table 1) with its performance on the pre-segmented transcriptions which were provided based on the actual room transition times. The difference can show the negative effect of inaccurate segmentation on the overall performance which is not increasing dramatically on transcriptions with more WER.

Table 1: *The performance ($P\%$) of the system on manual transcripts and ASR output of the clean and noisy speech data.*

Transcriptions	Map_1	Map_2	Map_3	Map_4	*Overall*
Manual	73.51	38.28	50.61	47.25	52.52
ASR, clean speech	69.47	29.13	42.68	38.80	45.15
ASR, noisy speech	67.91	27.19	41.11	37.10	43.46

Table 2: *The negative effect of segmentation error; a comparison between the overall performance of the system on the auto-segmented transcripts and its performance on the pre-segmented transcriptions using actual room transition times.*

Transcriptions	*Auto segmented*	*Pre-segmented*	*Negative effect*
Manual	52.52	54.49	1.97
ASR, clean speech	45.15	47.68	2.53
ASR, noisy speech	43.46	46.06	2.60

7 Conclusions

In this paper, we introduced a method for interpreting FRs' spoken description as a topological representation of the incident scene in a SAR scenario. Following the general principles of T-mapping, a novel landmark detection was introduced by framing this problem as a topic segmentation task on FRs' spoken reports. The presented method introduces new nodes to the map by identifying the changes in the content of spoken reports as an indication that the speaker has moved from one location to another, and text vectorization techniques are used to compute the correspondence between pairs of nodes and to estimate the T-map. The experiment results on manual and automatic transcriptions with different levels of WER confirm the low sensitivity of this method to highly imperfect speech recognition output. This speech-based T-map estimation system introduced a new source of information to the field of automatic map making. T-maps are not only suitable for human comprehension but also highly flexible to be integrated with other localization and mapping techniques such as SLAM (Simultaneous Localization and Mapping). It is anticipated that a careful integration of this system with other mapping techniques can provide a strong multimodal approach to this task.

Acknowledgments

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

References

[1] "Voice Radio Communications Guide for the Fire Service," Oct. 2008.

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

[3] C. Shimanski, "Situational Awareness in Search and Rescue Operations," *Mountain Rescue Association*, 2008.

[4] J. Boal, A. Sánchez-Miralles, and A. Arranz, "Topological simultaneous localization and mapping: a survey," *Robotica, Cambridge University Press*, vol. 32, pp. 803–821, 2014.

[5] Y.-T. Kuipers, Benjamin and Byun, "A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations," *Robotics and autonomous systems*, vol. 8, pp. 47–63, 1991.

[6] K. Lynch, *The image of the city*. MIT press, 1960, vol. 11.

[7] S. Mokaram and R. K. Moore, "Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario," in *Interspeech*, 2015.

[8] V. Claveau and S. Lefèvre, "Topic segmentation of TV-streams by watershed transform and vectorization," *Computer Speech & Language*, vol. 29, no. 1, pp. 63–80, Jan. 2015.

[9] E. Remolina and B. Kuipers, "Towards a general theory of topological maps," *Artificial Intelligence*, vol. 152, no. 1, pp. 47–104, 2004.

[10] J. Pradeep, V. and Medioni, G. and Weiland, "Visual loop closing using multi-resolution SIFT grids in metric-topological SLAM," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009, pp. 1438–1445.

[11] A. Ranganathan and F. Dellaert, "Online probabilistic topological mapping," *The International Journal of Robotics Research*, vol. 30, no. 6, pp. 755–771, 2011.

[12] M. Liu and R. Siegwart, "Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 310–324, 2013.

[13] I. Williams, B. and Klein, G. and Reid, "Real-Time SLAM Relocalisation," in *IEEE 11th International Conference on Computer Vision*, Oct. 2007, pp. 1–8.

[14] A. Angeli, D. Filliat, S. Doncieux, and J. A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[15] G. Youmans, "A New Tool for Discourse Analysis:The Vocabulary Management Profile," *Language*, vol. 67, pp. 763–789, 1991.

[16] B. J. Grosz and C. L. Sidner, "Attention, Intensions and the Structure of Discourse," *Computation Linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[17] J. Hirschberg and D. Litman, "Empirical studies on the disambiguation of cue phrases," *Computational linguistics*, vol. 19, no. 3, pp. 501–530, 1993.

[18] J. Hirschberg and C. Nakatani, "Acoustic Indicators of Topic Segmentation," in *Proceedings of the 5th International Conference on Spoken Language Processing ({ICSLP})*, 1998.

[19] M. a. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[20] M.-Y. Kan, J. L. Klavans, and K. R. McKeown, "Linear Segmentation and Segment Significance," *6th International Workshop of Very Large Corpora (WVLC-6)*, p. 9, 1998.

[21] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*. Association for Computational Linguistics, 2001, pp. 499–506.

[22] J. C. Reynar, "Topic segmentation: Algorithms and applications," *IRCS Technical Reports Series*, p. 66, 1998.

[23] G. Tur and R. De Mori, "Topic Segmentation," in *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, 2011, ch. 11, pp. 291–317.

[24] C. Guinaudeau, G. Gravier, and P. Sébillot, "Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations," in *Eleventh Annual Conference of the ISCA*, vol. 8, no. September, 2010, pp. 1365–1368.

[25] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–598, 1991.

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

[27] P. Lanchantin, P. Bell, M. Gales, and T. Hain, "Automatic Transcription of Multi-genre Media Archives," in *CEUR Workshop Proceedings Vol. 1012*, Marseille, France, 2013, pp. 26–31.