Speech-Based Location Estimation of First Responders in a Simulated Search and Rescue Scenario

Saeid Mokaram[1], *Roger K. Moore*[1]

[1]SpandH, Department of Computer Science, University of Sheffield, S1 4DP, United Kingdom

s.mokaram@sheffield.ac.uk, r.k.moore@sheffield.ac.uk

Abstract

In our research, we explore possible solutions for extracting valuable information about first responders' (FR) location from speech communication channels during crisis response. Fine-grained identification of fundamental units of meaning (e. g. sentences, named entities and dialogue acts) is sensitive to high error rate in automatic transcriptions of noisy speech. However, looking from a topic-based perspective and utilizing text vectorization techniques such as Latent Dirichlet Allocation (LDA) make this more robust to such errors. In this paper, the location estimation problem is framed as a topic segmentation task on FRs' spoken reports about their observations and actions. Identifying the changes in the content of a report over time is an indication that the speaker has moved from one particular location to another. This provides an estimation about the location of the speaker. A goal-oriented human/human conversational speech corpus was collected based on an abstract communication model between FR and task leader during a search process in a simulation environment. Results show the effectiveness of a topic-based approach and especially low sensitivity of the LDA-based method to the highly imperfect automatic transcriptions.

Index Terms: spoken language understanding, speech recognition, topic segmentation, Latent Dirichlet Allocation (LDA), human/human conversation

1. Introduction

Speech is the single most important source of situational information during crisis response. It is widely used for transferring critical information about the location of First Responders (FR) and their ambient conditions to the Task Leader (TL) [1]. Accurate assessment of the location of FR is known to be one of the main enhancing factors for situation awareness and more successful missions within the Search and Rescue (SAR) context [2]. Despite the importance of such location information, today, assimilation of such information and key elements from speech is almost entirely done manually. Automatic extraction of this rich source of information for location estimation reduces the risk of human related errors in large and fast moving SAR operations. The importance of location estimation based on speech communications has been envisaged in the observational-speech-system [3]. Yet, technical difficulties such as high word error rate (WER) in automatic speech recognition (ASR) transcripts and understanding spontaneous human/human spoken language communications present major challenges for implementing such system. To the best of our knowledge, this paper presents the first attempt in providing a speech-based location estimation approach based on the speech communication transcripts in a simulated urban search and rescue (USAR) scenario.

Although current ASR systems are capable of achieving high recognition performance [4], they still have a lot of difficulties in highly noisy environments. Fine-grained identification of the fundamental units of meaning (e. g. sentences, named entities and dialogue acts) is sensitive to high WER in the automatic transcriptions [5]. In contrast, looking from a topic-based perspective and utilizing state-of-the-art in text vectorization techniques would result in more robust systems to such errors [5, 6].

The explanations of FRs about their observations, actions and events are highly associated with their location. Consequently, the information content of their spoken reports change while they are exploring the incident scene. Significant changes in content would be an indication that they have moved from one particular location to another. FRs also tend to signal their TL about their intention of moving from one particular place (room) to another. As Cassell et al. mentioned [7], topic changes may correspond to changes in physical posture of speaker. So, identifying these changes and signals results in segmenting a long speech report into short coherent units which each unit is representing a particular visited location. This can provide an estimation of the location of the speaker. The similarity between this task with the topic segmentation and identification problem makes it possible to apply the state-of-the-art in this field.

Vectorization techniques such as Latent Dirichlet Allocation (LDA) [8] and Term Frequency-Inverse Document Frequency (TF-IDF) [9] were applied in order to describe utterances in the vector space model (VSM). To identify the transition signals and changes in the content of spoken reports, the similarity between sequence of utterances and number of pivot documents was computed in the VSM. Using the simplified immersion paradigm in watershed technique for a 1D signal which is introduced in [10] topic breaks are identified on both LDA and TF-IDF representations. The content of each segment was then compared against numbers of pivot documents which each represent a specific location. Merging adjacent segments which are identified as the same location reduced the over-segmentation problem. The high quality of segmentation (≈ 0.13 WindowDiff error rate) shows the effectiveness of topic-based approaches in this task. Experiments on the ASR transcripts of both clean speech (with 32.4% WER) and noisy speech (with 41.6% WER) show that the LDA-based approach is less sensitive to this increase of WER in compare to the TF-IDF method.

In the remainder of this paper, a brief summary of the principle of segmentation for spoken document is presented in the next section. Then in Section 3, the Sheffield Search and Rescue (SSAR) corpus as a goal-oriented two-party human/human conversational speech corpus in the context of USAR scenario

is introduced. The topic-based approach for estimating FR location based on voice communications is explained in Section 4. Then the results are presented in Section 5 followed by discussion and conclusions.

2. 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 [11]. 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 tend to signal to the audience about switching from one topic to another by using various words/phrases or prosodic cues [12, 13, 14]. 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 compare to content-based techniques.

Different approaches have been introduced both for content-based and boundary-based. The TEXT-TILING system [15] 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 [16] is defined as finding the boundaries on a representation of text as weighted lexical chains. Utiyama and Isahara [17] applied a Hidden Markov Model (HMM) based statistical approach to measure lexical cohesion with the help of language modelling. DotPlotting [18] and C99 [19] both 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.

2.1. Vectorization for segmentation

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 [5]. In order to deal with short segments with very few common words, Guinaudeau et al [20] integrated the semantically related terms to the HMM segmentation model in [17] to extend the description of the possible segments. Claveau and Lefevre [21] used vectorization techniques such as LDA 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 [6, 21]. They also adopted the watershed transform which is a famous morphological method for image segmentation [22] and achieved a high quality of segmentation.

2.2. Evaluation Metric

The recall and precision metric has often been applied to the topic identification problem. 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 RP can be inconsistent without any preprocessing, Beeferman et al. [23] proposed the Pk-score, which has been widely used. However, it has been shown that the Pk-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 compare to false alarms and missing boundaries; 3) not detecting new segments with size smaller than k; 4) not clear meaning of (cannot be interpreted as an error percentage) [24]. Based on that, WindowDiff (WD) has been proposed [24] which is usually preferred for evaluating segmentation systems. It can be seen as an error rate, which lower WD scores indicate better segmentation accuracy. It is defined as:

$$WD(r,h) = \frac{1}{N-kj} \sum_{i} \mid b(r_i, r_{i+k}) - b(h_i, h_{i+k}) \mid \quad (1)$$

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

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

WD-scores and RP are adopted in this paper for explaining the performance of transition detection and location identification respectively.

3. SSAR: a human-human spoken conversation corpus

A goal-oriented two-party human/human conversational speech 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 setup is visualized in Figure 1. In this arrangement, FR and TL were located in separate room. 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 what kind of room the FR was in at each time. Inspired from simulation training systems which are being used by some fire departments, a simulated indoor environment was designed. Four different settings for the simulated environment was designed in order to have multiple levels of complexity and difficulty. Figure 2 shows the top-view of the Map_4 settings. 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 similar type of rooms; for example Map_2 has two different bedrooms. In total 13 different types of indoor locations ($RoomTypes$) such as *kitchen, bedroom, computer-lab, etc.* 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.

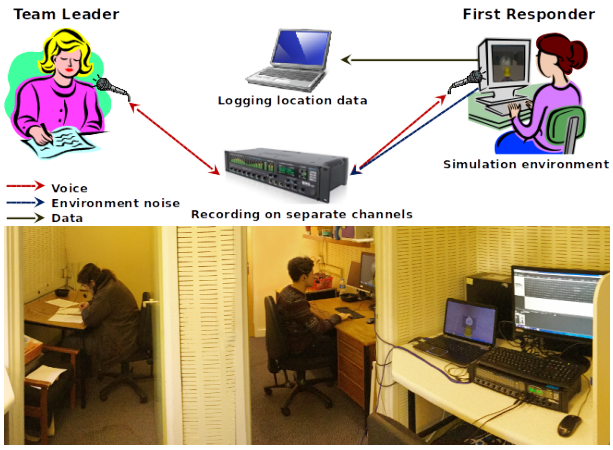

Figure 1: *top: the recording scenario, bottom: the recording set-up in two separate quiet rooms.*

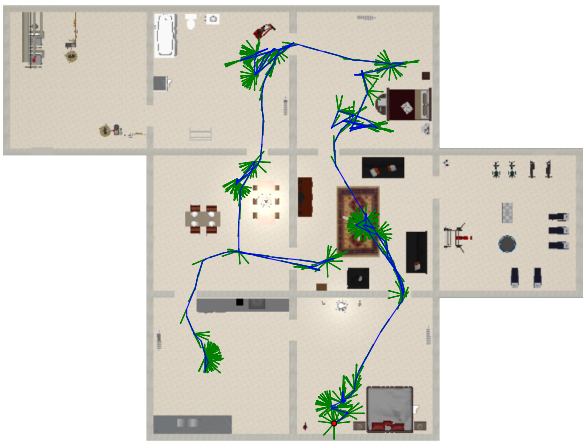

Figure 2: *An example of a map-setting which was used in data collection. The motion trajectory of a participant is plotted on the* Map_4.

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.

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. In all experiments, the location of FR was correctly estimated by the participants who played the role of TL. This confirms that the amount of exchanged information through voice channel is sufficient for a human subject to estimate the location of speaker. Each recording has an average length of ≈ 7.25 minutes. The corpus contains 12 hours of conversational speech with word level manual annotations.

4. Speech-Based Location Estimation

4.1. Transition detection

In the SAR context, the FR tends to signal to the TL while moving from one location to another by using various words and phrases such as, "*now*", "*going to go*", "*this room*" and so forth. Since it is more likely that the adjacent utterances of transitions contain these kind of signals, K utterances before and after each transition were considered as transition-related utterances (best result from $K = 2$). 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 data. 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 LDA and TF-IDF vectorization principle, both window and TPD were projected into the VSM and the cosine similarity between their vectors were computed. An example of this LDA/cosine similarity between the vectors is visualized in Figure 3 (top). The computation of the LDA/cosine distance (or similarly for the TF-IDF/cosine) can be written as:

$$D(i) = cosine\Big(LDA(TPD), LDA(u_{i-k,i+k}) \Big) \qquad (3)$$

Based on the explained watershed-based segmentation approach in [10], an estimation of the transition times were calculated. While applying a small window ($w = 5$) increases the time-precision in transition detection, it also increases the chance of false-alarms and over-segmentation. An example of this over-segmentation can be found at utterance 60 in Figure 3 (top).

4.2. Location identification

The segmentation process provides estimations about the transition times. These estimations divide a long sequence of utterances into smaller sections. In a correct segmentation, the adjacent sections must be related to different location. The next step in location estimation is to identify and label these segments as locations. To this purpose, the similar vectorization principal in the segmentation task was used with a difference that here the entire section was compared against 13 room-pivot-documents (RPD). Each of these RPDs was built by combining all the utterances related to one of the 13 types of locations in the training data. Comparing each segment as a whole resulted in more accurate identifications in contrast to single utterance or a small window comparison. This is because a larger set of utterances contains more information and is more robust to the added noise by the ASR module. The probability distribution of 13 classes for each segment was computed by normalizing the segment similarities to the RPDs. As a result, each utterance in every section was labelled based on its most likely $RoomType$. Figure 3 (bottom) is visualizing an example of this probability distribution which is computed using LDA/cosine on a conversation. For example, it is clear that the first segment is more related to the $RoomType_1$ than the rest so, it is labelled as $R1$. Finally, relying on the high accuracy of this location identification, two adjacent segments which are classified as a same location was merged in order to reduce the over-segmentation of transition detection step. As an example, this can be seen in 5th and 6th estimated segments in Figure 3.

4.3. Experiments

Two sets of experiments were conducted on clean speech data and speech with the background environment noise. The ASR system used for the experiments was accessed through webASR [25]. The specific system used was a 2-pass DNN-GMM-HMM tandem system trained on 95 hours of speech from 327 British television and radio broadcasts. The language model used was a 3-gram based on the interpolation of multiple language models trained on meeting transcripts, broadcast subtitles and telephone

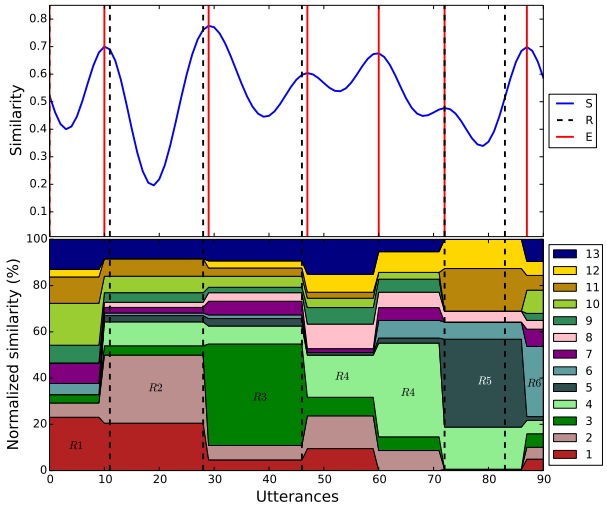

Figure 3: *top: Shows the LDA/cosine similarity (S) between utterances and TPD. Here a sliding window with the size of* 5 *was used. The reference (R) and estimated (E) transition lines are plotted; bottom: Shows the probability distribution of* 13 $RoomTypes$ *for each segment. In this example a participant has visited rooms in the following order: R1→R2→R3→R4→R5→R6.*

conversations, 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 [26], 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. The introduced segmentation and identification process were applied on ASR transcripts of both noisy and clean speech data. The K-Fold cross-validation ($k = 10$) was used in order to divide the data into train-dataset and validation-dataset. Document vectors were produced by applying LDA (number of topics = 30) and TF-IDF scheme and then the similarity between two vectors were measured by computing their cosine similarity distance.

5. Results and discussion

WD-score was used as the quality measure for the transition detection and recall (R), precision (P) and F1-measure (F1) was used to measure the location identification performance. Table 1 shows the results obtained by the transition detection step using both LDA and TF-IDF methods on the clean and noisy speech data. It is notable that the LDA-based approach has lower WD error on both clean and noisy data in compare to TF-IDF. The WD results also shows that in spite of a considerable increase in the WER (9.2%), both systems did not receive a high negative impact from that and their WD error did not increase dramatically. In the presence of high transcription WER, the LDA-based method in particular, performed much better in compare to the TF-IDF. This indicates the robustness of this system to such errors.

Table 2 shows results after applying the identification and merging step. Here R, P, and F1 are reflecting the identification performance, and WD presents the quality of segmentation after merging. The results in this table shows a general reduction in

Table 1: The WD error of LDA/cosine and TF-IDF/cosine approaches on the clean and noisy speech transcripts with 32.4% and 41.6% WER respectively.

Methods	Speech data	WD
LDA	clean	0.212
	noisy	0.219
TF-IDF	clean	0.303
	noisy	0.319

WD in compare to the transition detection performance in Table 1. This is an indication of the positive effect of merging on the transition detection performance. It is also notable that, an increase of 9.2% WER in the transcripts errors, leads into just 0.006 WD increase in the error of LDA-based method which is 60% better than 0.015 WD increase in TF-IDF approach. This again confirms the advantage of using LDA over TF-IDF approach.

Table 2: *The performance of LDA/cosine and TF-IDF/cosine room identification (R, P and F1); and the WD-score of the segmentation approaches after the merging step.*

Methods	Speech data	R	P	F1	WD
LDA	clean	67.30	75.12	70.99	0.136
	noisy	65.90	70.10	67.94	0.142
TF-IDF	clean	51.16	49.53	50.33	0.238
	noisy	48.33	44.94	46.58	0.253

It is important to note that, since R,P and F1 were calculated at the utterance level, they reflect both identification and segmentation performance together; therefore, they can be considered as the overall quality of location estimation. The high quality of the presented results in the segmentation task is comparable with the results presented in [10].

6. Conclusions

In this paper, we introduced a speech-based localization system for estimating the location of FRs based on their speech communications in a USAR scenario. The location estimation problem was framed as a topic segmentation/identification task on FRs' spoken reports about their observations and actions. As a result, tracking the changes in the content of their spoken reports over time provided an estimation about the location of the FR in a simulated environment. The LDA vectorization technique made this possible to match text segments that do not share common words. This enabled it not only to perform better than the TF-IDF approach, but also being more robust to errors in highly imperfect automatic transcriptions. The experiment results confirmed this by showing that, an increase of 9.2% WER in the transcripts errors, leads into just 0.006 WD increase in the error of LDA-based method which is 60% better than 0.015 WD increase in TF-IDF approach. This speech-based location estimation system introduced a new source of information to the field of localization. It is anticipated that a careful integration of this system with other localization techniques such as SLAM (Simultaneous Localization and Mapping) can provide a strong multimodal approach for the location estimation task.

7. Acknowledgements

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

8. References

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

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

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

[4] M. Ostendorf, B. Favre, R. Grishman, D. Hakkani-Tür, M. Harper, D. Hillard, J. Hirschberg, H. Ji, J. G. Kahn, Y. Liu, S. Maskey, E. Matusov, H. Ney, A. Rosenberg, E. Shriberg, W. Wang, and C. Wooters, "Speech segmentation and spoken document processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 59–69, May 2008.

[5] 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.

[6] M. Morchid, M. Bouallegue, R. Dufour, G. Linares, D. Matrouf, and R. De Mori, "I-vector based Representation of Highly Imperfect Automatic Transcriptions," in *Interspeech*, no. September, 2014, pp. 1870–1874.

[7] J. Cassell, Y. Nakano, T. Bickmore, C. Sidner, and C. Rich, "Non-Verbal Cues for Discourse Structure," *Thirty-ninth Annual Meeting of the Association of Computational Linguistics*, pp. 106–115, 2001.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[9] G. Salton, "A Theory of Indexing, regional Conference Series in Applied Mathematics," *Society for Industrial and Applied Mathematics*, no. 18, 1975.

[10] 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.

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

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

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

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

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

[16] 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.

[17] 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.

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

[19] F. Y. Y. Choi, "Advances in domain independent linear text segmentation," in *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*. USA: Association for Computational Linguistics, 2000, p. 8.

[20] 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 International Speech Communication Association*, vol. 8, no. September, 2010, pp. 1365–1368.

[21] V. Claveau and S. Lefèver, "Topic segmentation of TV-streams by mathematical morphology and vectorization," *Interspeech*, pp. 1105–1108, 2011.

[22] 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.

[23] D. Beeferman, A. Berger, and J. Lafferty, "Statistical models for text segmentation," *Ml*, vol. 210, no. 1-3, pp. 177–210, 1999.

[24] L. Pevzner and M. a. Hearst, "A Critique and Improvement of an Evaluation Metric for Text Segmentation," *Computational Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.

[25] 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.

[26] 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.