The non-Verbal Structure of Patient Case Discussions in Multidisciplinary Medical Team Meetings

Saturnino Luz, Trinity College Dublin

Meeting analysis has a long theoretical tradition in social psychology, with established practical ramifications in computer science, especially in computer supported cooperative work. More recently, a good deal of research has focused on the issues of indexing and browsing multimedia records of meetings. Most research in this area, however, is still based on data collected in laboratories, under somewhat artificial conditions. This paper presents an analysis of the discourse structure and spontaneous interactions at real-life multidisciplinary medical team meetings held as part of the work routine in a major hospital. It is hypothesised that the conversational structure of these meetings, as indicated by sequencing and duration of vocalisations, enables segmentation into individual patient case discussions. The task of segmenting audio-visual records of multidisciplinary medical team meetings is described as a topic segmentation task, and a method for automatic segmentation is proposed. An empirical evaluation based on hand labelled data is presented which determines the optimal length of vocalisation sequences for segmentation, and establishes the competitiveness of the method with approaches based on more complex knowledge sources. The effectiveness of Bayesian classification as a segmentation method, and its applicability to meeting segmentation in other domains are discussed.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems; I.5.2 [**Pattern Recognition**]: Classifier design and evaluation; H.5.3 [**Group and Organization Interfaces**]: Computer-supported cooperative work

General Terms: Human Factors

Additional Key Words and Phrases: Search of spontaneous speech, meeting analysis, dialogue segmentation, multidisciplinary medical team meetings, audio analysis

ACM Reference Format:
Luz, S. 2012. The non-Verbal Structure of Recorded Multidisciplinary Medical Team Meetings. ACM Trans. Inf. Syst. 30, 3, Article 19 (June 2012), 24 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

Group dialogue at meetings has been a topic of systematic study from quantitative and qualitative perspectives since at least the 50's, with the works of Bales [1950] and others. This line of work has investigated issues such as group performance [McGrath 1991], group cohesiveness, and the process of verbal and non-verbal activities [Hackman and Morris 1975; Dabbs and Ruback 1987]. Advances in computer technology have stimulated research on similar topics in the computer science disciplines of human-computer interaction (HCI) and computer supported cooperative work (CSCW). This includes studies of video-mediated meetings [Olson et al. 1993;

Author's address: S. Luz, School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, Ireland.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

Finn et al. 1997], as well as the development of meeting support systems [Galegher and Kraut 1994; Olson et al. 1993; Gutwin and Greenberg 1999], and tools to support the capture of conversations [Hindus et al. 1993] and meeting records [Richter et al. 2001].

Capturing, coding and analysing meetings have formed part of social psychology and CSCW group interaction research from early days. More recently, as multimedia recording of face-to-face and remote meetings has become technically and economically viable, HCI and language technology researchers have argued for the development of systems to support indexing, searching and browsing of recorded multimedia meeting data [Banerjee et al. 2005; Waibel et al. 2001; Kazman et al. 1996]. The topic of "meeting browsing" [Bouamrane and Luz 2007; Tucker and Whittaker 2005], in particular, has received considerable attention from researchers from a variety of backgrounds. Large research projects, such as the ICSI Meeting Recorder [Janin et al. 2003], AMI/AMIDA [Renals et al. 2007], ISL Meeting Room [Burger et al. 2002], M4 [McCowan et al. 2005], VACE-II [Chen et al. 2006] and the NIST Meeting Corpus [Garofolo et al. 2004], have helped lay the foundations for automatic analysis of meeting contents by collecting large, extensively annotated corpora of meeting data, which in turn led to the development of a number of techniques for search and visualisation of meeting data [Waibel et al. 2001].

However, despite of its eminently practical motivations, research on meeting browsers shared with its predecessors in social psychology the same reliance on data obtained under laboratory conditions which has been criticised by McGrath [1991] and others, in the group research literature. On the one hand, fieldwork and qualitative research have been carried out on how people create records of real meetings through minute taking and personal notes [Whittaker et al. 2006] and by technology enhanced means [Moran et al. 1997]. On the other hand, the most widely used data sets available to technology developers consist mainly of scenario driven meetings [Renals et al. 2007; McCowan et al. 2003; Bouamrane and Luz 2007] and meetings among the researchers themselves [Janin et al. 2003]. While, as Carletta [2007] argues, the relative homogeneity of the data obtained in controlled environments facilitates intrinsic evaluation of the different machine learning and natural language processing techniques commonly employed in the analysis of meetings, questions remain as to how effective these techniques can be in more realistic application scenarios.

These questions can only be answered, of course, on a case by case basis, in the context of system development. Observation and data collection in real workplace environments can nevertheless help assess the applicability of these techniques to different kinds of meeting data. This paper reports on one such assessment applied to audio and video data collected as part of a three-year ethnographic study of a multidisciplinary medical team and their regular meetings [Kane and Luz 2006; Kane 2008], and subsequently annotated by the researchers for vocalisation event durations and speaker identity. Multidisciplinary medical team meetings (MDTMs) are meetings in which several specialists gather in order to discuss patient cases, agree on a diagnosis, and make treatment and patient management decisions. The technique in question is segmentation of conversations into topics, in a broadly defined sense [Galley et al. 2003; Banerjee et al. 2005; Hsueh et al. 2006; Dielmann and Renals 2007].

A typical MDTM lasts over one hour and consists of a sequence of patient case discussions (PCDs). We identify each of these PCDs with "topics", in the sense that they have well defined conceptual boundaries and can be categorised into different types, such as medical and surgical discussions, local patient and referral patients, co-located PCDs and remote PCDs, etc [Luz and Kane 2009]. Current approaches to meeting topic segmentation employ combinations of feature sources, including lexical features (or "bags of words") obtained from the output of a speech recogniser, conversational fea-

tures (lexical cohesion statistics as well as dialogue structure, vocalisation and silence statistics) [Galley et al. 2003], prosodic features [Shriberg et al. 2000], video features [Dielmann and Renals 2007; Hsueh and Moore 2007], and other contextual features such as dialogue type and speaker role [Hsueh and Moore 2007]. Only a few of these information sources can be reliably extracted from recordings obtained at a real MDTM, where the fast pace of the dialogue, the large number of participants, the diverse composition of the medical teams, and other factors make clean recording of individual speakers a practical challenge. Very high word error rates for automatic speech recognition, for instance, would preclude the use of dialogue acts, lexical features and lexical cohesion statistics for our MDTM data (even though some topic segmentation systems have been shown to be resilient to moderate word error rates [Garofolo et al. 1999; Hsueh and Moore 2007] in other domains). This is, partly, our motivation for investigating the use of "content-free" features of talk for segmentation[1].

However, the investigation is also motivated by theoretical interests. Extending early work on Markovian models for dyadic interactions and monologues [Jaffe and Feldstein 1970], Dabbs and Ruback [1987] have argued that such content-free features as patterns of turn-taking (vocalisation) and silence can tell an analyst much about the nature and structure of a meeting. In the case of MDTMs (and patient case discussions, in particular), despite being fast-paced and apparently chaotic to an outside observer, the conversations are highly structured events where the participants have very well defined roles, according with their medical specialities, which determine to a great extent their patterns of participation in the meeting.

As a matter of practical relevance to both indexing of meeting content for browsing and information retrieval purposes and the theoretical analysis of meeting process, this paper investigates whether segmentation of MDTMs into their constituent PCDs can be reliably performed based on speaker roles and patterns of vocalisation and silence. These features form part of what is arguably the simplest account of the sequential structure of dialogue [Dabbs and Ruback 1987] and therefore seem like a promising starting point, from which analyses incorporating richer elements (such as transcription, semantic interpretation and visual modalities) can be further developed.

The paper is structured as follows. The next section outlines the background of this study, describing the PCD and the MDTM in greater detail, presenting basic descriptive statistics of participant roles and interaction, and motivating the research from a practical point of view. Section 3 presents a review of related work on topic segmentation, highlighting the similarities and differences between MDTM segmentation and meeting topic segmentation in general, and tracing back the origins of many dialogue segmentation approaches to early work on text segmentation. This is followed by the presentation of our approach to data representation, the data preparation and annotation procedures adopted for this study. The main segmentation technique is then introduced, followed by the results of a cross validation experiment performed in order to assess the effectiveness of combining Naïve Bayes classification and the proposed content-free representation in detecting PCD boundaries. The analysis of results is complemented by a baseline analysis, a study of the effect of diarisation errors on segmentation accuracy and an analysis of the effect of redundancy on the Naïve Bayes classifier. Comparisons with hidden Markov models, decision trees, kernel methods and nearest neighbour classifiers are presented, along with a discussion of evaluation issues, and other state of the art approaches to topic segmentation. Conclusions, and plans for future work close the paper.

[1]Incidentally, these issues would also motivate the investigation of video features (e.g. communicative gestures, or use of presentation aids) which are not addressed in this paper.

Fig. 1. A MDTM in a tertiary referral hospital.

2. BACKGROUND

The meetings analysed in this paper take place in a hospital setting and are attended by a varying number of participants who constitute a multidisciplinary medical team. Multidisciplinary medical team meetings are an established part of the process of diagnosis and treatment of cancer patients, and are a practice recommended by several national health services [Calman-Hine 1995]. In an MDTM, health professionals of different specialities meet to discuss diagnosis, treatment options and patient management. Additionally, these meetings serve educational purposes (training of students and junior doctors) and broader healthcare management and organisational functions [Kane 2008].

The MDTM is structured as a sequence of PCDs, where the patient's medical record is reviewed, evidence from pathology and radiology is presented, the possibility of surgery is discussed, and a patient management plan is agreed. In addition to the medical team, the meetings are generally attended by medical students and junior staff, who do not play an active role in the discussions. The presentations and discussions make intensive use of visual aids (e.g. display of pathology slides on a large screen, radiology images on high-resolution displays, etc), and are often also attended by remote participants connected through teleconferencing. Support for collaboration at MDTMs is a topic that has attracted the interest of the CSCW community lately, and detailed analyses of organisational processes surrounding the meeting, its different functions in the hospital environment, and mechanisms that add dependability to their decision making have been conducted [Robertson et al. 2010; Groth et al. 2009; Kane and Luz 2006].

Figure 1 shows the physical environment in which the MDTMs recorded for the corpus used in this study take place. It is a dedicated teleconferencing room equipped with projection equipment, a high resolution screen for radiological images, a large plasma screen, as well as microscopes and document readers which can be connected to the large display. The recordings were taken from two separate sources: (a) the existing teleconferencing equipment fitted into the meeting room, which recorded the audio through a pressure-zone microphone and alternated recording of the video chan-

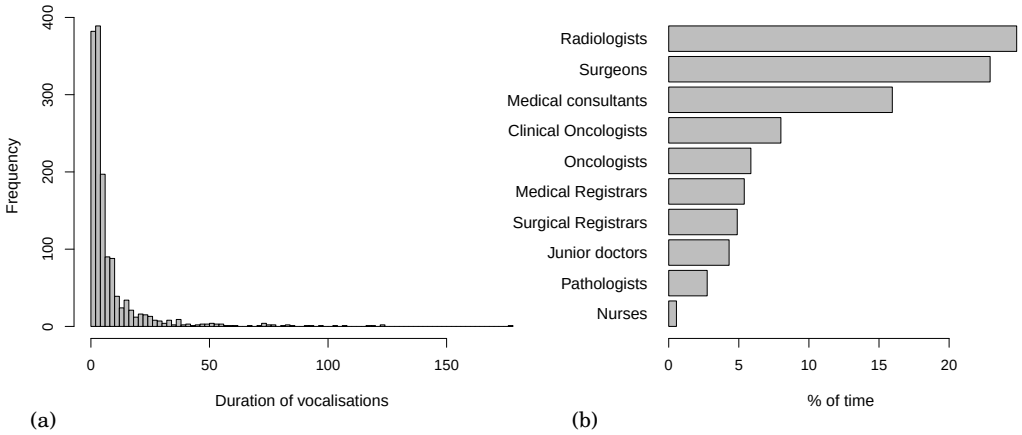

Fig. 2. Amount of talk in MDTMs: (a) according to duration of individual vocalisations and (b) distributed by medical roles.

nel between a view of the participants and views of the different medical images under discussion, and (b) a high-end camcorder mounted on a tripod which recorded the audio through a sensitive directional microphone. These two sources were aligned (synchronised) using a multimedia annotation tool. While small and lacking in annotation detail if compared to the major meeting corpora mentioned above (Section 2), the MDTM corpus is unique in that it was collected in a real-world environment, with the meeting participants engaged in a complex professional task [Kane 2008].

Over 28 hours of meeting data were collected, in total. For the study reported in this paper, a data set of 54 PCDs (approximately 220 minutes) were segmented and annotated using the ELAN Linguistic Annotator [MPI 2005]. The original purpose of data collection was to investigate the diagnosis and decision making processes of multidisciplinary medical teams [Kane and Luz 2006; 2009] within an interaction analysis framework [Jordan and Henderson 1995]. This initial work revealed, among other things, that although the medical team works under severe time constraints and consequently the PCDs need to be well structured, the group managed to balance task and socio-emotional exchanges, which as McGrath [1991] suggests, is a means of avoiding tension and negative reactions in group collaboration.

The distribution of talk at MDTMs is very skewed, as evidenced in Figure 2(a), which shows that the vast majority of vocalisations are of short duration. This is in agreement with the general pattern in multi-party dialogues [Dabbs and Ruback 1987]. MDTM vocalisations tend to last longer than those in scenario based meetings such as the ones recorded for the AMI corpus [Carletta 2007]. The mean duration of an MDTM vocalisation is 8.2s (median 3.5s) while the a mean duration of an AMI meeting vocalisation is 3.9s (median 1.4s) [Luz and Su 2010]. The AVERAGE number of (active) participants in an MDTM is also greater, as is the number of distinct roles they perform in meetings. We identified 10 distinct participant roles, whose proportional contributions (in terms of amount of talk) to the meetings is summarised in Figure 2(b). This again is in contrast to the AMI data where only 4 roles are defined: user interface designer, industrial designer, project manager and marketing expert. Unlike the speakers in the broadcast news data analysed by Vinciarelli [2007], who can play different roles at different times during the broadcast, each speaker in the MDTM corpus has a unique medical specialist role throughout the PCDs. The mapping from speaker identities to

Table I. Descriptive statistics (mean and standard deviation) for patient case discussions in the MDTM corpus

Description	mean	SD
Mean vocalisation length (per PCD)	8.55 sec.	4.45
Number of speakers per PCD	8.15	1.8
Vocalisations per PCD	29.6	17.6
Number of vocalisations per speaker (per PCD)	3.507	1.66
Vocalisations per minute	8.76	4.0
Group Vocalisation length	1.74 sec.	1.85
Silence per PCD	4.12%	3.7
Participation ratio	0.39	0.25
Entropy (speaker transitions)	0.78	0.35
Entropy (vocalisations lengths)	2.23	0.46

roles is not one-to-one, however, as more than one speaker can perform the same role in the same PCD.

A total of 21 different speakers actively participated in the MDTMs. The remaining time is distributed between pauses (silences, 3.4%) and "group vocalisation" (overlapping speech, 1.2%). Averaged with respect to PCDs (Table I) talk duration is consistent with the general MDTM figures. Table I also shows the average number of vocalisations per participant during a PCD, the average number of speakers participating in each case discussion and the average duration of intervals of silence. The last row contains the mean values for a metric we call *participation ratio* [Luz and Kane 2009]. The participation ratio of a meeting attendee is defined as the ratio of the number of PCDs in which the attendee took active part to the total number of cases discussed. The figures for mean participation ratio (over n speakers) in Table I were calculated as $r = \sum_i^n |\mathcal{C}_i|/(n|\mathcal{C}|)$, where $\mathcal{C}_i$ represent the set of PCDs in which speaker s_i produced at least one vocalisation and $\mathcal{C}$ is the entire set of PCDs.

Participation ratio figures summarise variability in the composition of the groups across case discussions. Table I indicates a high degree of variation, showing that a speaker will on average take part in only around 39% of all PCDs. A different measure of variability is given by the Entropy (H) for the probability distribution of vocalisations by n speakers, calculated by averaging over the probabilities p_i that speaker s_i is speaking at a given time during PCD, in the usual way: $H = \sum_i^n p_i \log 1/p_i$. The H score for speaker transitions ($H = 0.78$, $sd = 0.35$) reveals a predictable pattern of speaker transitions while the entropy of the vocalisation length indicates a process that is less predictable ($H = 2.23$, $sd = 0.46$), though the amount of uncertainty in the distribution of vocalisation length is still quite small considering that the we have on average more than eight participants in a PCD.

2.1. The practical relevance of MDTM segmentation

It has been acknowledged in group research that descriptive statistics such as the ones shown in Table I alone do not suffice to characterise a meeting. The interaction process that links group task inputs to task outputs also needs to be considered [Hackman and Morris 1975; Dabbs and Ruback 1987]. The Interaction Process Analysis proposed by Bales [1950] and related systems provide an account of process which is based on careful coding of ongoing group interactions. Dabbs and Ruback [1987] argue that, although useful for analysis, the method of coding the content of speech interaction with reference to a system of categories tends to miss important information found in the more general paralinguistic features of meetings. From a CSCW perspective, a system such as the one outlined in this paper, capable of automating the collection of content-free (paralinguistic) features and segmenting the recorded data into meaningful subunits would provide a tool for analysing meetings with respect to their effectiveness, the impact of new meeting-support technologies on the interaction, etc.

In terms of meeting indexing, searching and browsing, structuring a multimedia meeting record by topics of discussion could help users access audio content even in the absence of speech transcription by providing reference points on the time line [Bouamrane and Luz 2007; Moran et al. 1997]. While a variety of techniques have been employed to add structure to different kinds of audio recordings, including role identification on radio broadcasts [Barzilay et al. 2000], summarisation of broadcast news based solely on prosodic features [Maskey and Hirschberg 2006] and "rich transcription" of meetings [Fiscus et al. 2008], the approach presented in this paper aims to identify reference points on meeting recordings based exclusively on the amount and structure of talk. While it is clear that a complete meeting storage and retrieval system will also require modality translation algorithms such as speech to text conversion and video analysis, as currently pursued by many researchers, we believe that a focus on content-free features will contribute in its own right to a better understanding of the organisation of meeting data.

3. BRIEF REVIEW OF DISCOURSE SEGMENTATION AND RELATED WORK

For the purposes of meeting segmentation PCDs can be regarded as sequences of vocalisations grouped under a common topic (i.e. the discussion of a particular patient's case). In this sense, the task of segmenting MDTMs into PCDs is similar to the topic segmentation task as defined by Galley et al. [2003] and tackled in recent work on meeting analysis [Dielmann and Renals 2007; Hsueh 2008; Hsueh and Moore 2007; Banerjee and Rudnicky 2007].

Meeting segmentation has been influenced by early work on text segmentation [Hearst 1997; Beeferman et al. 1999], with which it shares evaluation metrics and methods. Approaches to broadcast news segmentation [Rosenberg et al. 2007; Shriberg et al. 2000] and lecture segmentation [Malioutov et al. 2007] have also influenced meeting segmentation research. However, it is generally acknowledged that segmentation of spontaneous speech produced by interacting speakers in a group is more challenging than text segmentation, where the topic structured is (in most cases) carefully designed by the writer [Gruenstein et al. 2005], and segmentation of broadcast news audio and other non-conversational speech, where the production environment and other contextual factors might provide acoustic clues as to where segment boundaries lie [Rosenberg et al. 2007; Gruenstein et al. 2005].

Related techniques are the identification and clustering of individual group actions [Zhang et al. 2006; McCowan et al. 2005; McCowan et al. 2003] and the labelling of topics [Blei and Moreno 2001]. While the present work is not concerned with these tasks, we acknowledge that they could play an important role in the automatic structuring of spontaneous speech. We investigated labelling issues (PCD content categorisation) elsewhere [Luz and Kane 2009]. Here, the meeting is simply treated as a sequence of vocalisations and pauses, and an attempt is made to mark out those vocalisations which signal the beginning of a PCD.

Boundary vocalisations are similarly distributed for PCDs in our MDTM corpus, where only about 3.6% of all vocalisations indicate the start of a PCD, and in the AMI corpus, where about 3.3% of talk spurts indicate a topic change [Hsueh and Moore 2007]. However, despite these similarities MDTM segmentation differs from meeting topic segmentation in that the latter seeks to identify segments that are different as they appear in the vocalisation sequence, whereas the former aims to segment the stream into essentially similar sub-sequences. Topics in the AMI corpus, for instance, can be categorised as "top-level" and "functional topics"[2] [Hsueh and Moore 2007] denoting segments that could also be described as "meeting states" [Banerjee and Rud-

[2]http://corpus.amiproject.org/documentations/annotations

nicky 2004], such as “presentation”, “discussion”, “opening”, “closing”, “agenda” etc, which can then be subdivided into sub-topics, forming a shallow hierarchy which is usually flattened for the purposes of segmentation. Similarly, Gruenstein et al. [2005] annotated the ICSI corpus [Janin et al. 2003] hierarchically according to topic, identifying, in addition, action items and decision points. For some applications, however, annotation focuses simply on topic changes that produce high inter-annotator agreement scores, with no further specification of topic label or discourse structure [Galley et al. 2003].

In MDTM segmentation, due to the self-contained nature of PCDs, annotators have little difficulty in identifying case discussion boundaries. The consistency of the manual segmentation of the MDTM corpus was ensured by the close collaboration between the researcher who gathered the data and members of the medical team who reviewed the annotation. It should also be remarked that MDTM segmentation can also be hierarchical, since PCDs exhibit an identifiable set of internal discussion states, including presentation of symptoms and clinical findings, questioning and correlation of pathology, radiology and examination data, disease stage classification, discussion of patient management options, and articulation of the decision agreed [Kane and Luz 2009]. However, this level of topic structure has not been fully annotated in the MDTM corpus and is therefore not addressed in this paper.

Different strategies have been employed for conversational topic segmentation. As mentioned above, Galley et al. [2003] model meeting topic segmentation after a text segmentation approach (namely, TextTiling [Hearst 1997]), relying on transcribed speech to compute lexical cohesion probabilities for adjacent analysis windows. Renals and Ellis [2003], on the other hand, consider “non-lexical methods” for segmentation which bear some similarity with our approach in that their data representation is based on patterns of talk spurts encoded as transition matrices. However, their segmentation algorithm, which is analogous to acoustic speaker segmentation using the Bayesian Information Criterion, does not produce satisfactory results, leading the authors to speculate that “turn pattern boundaries are not directly related to discussion topics” [Renals and Ellis 2003]. The results presented in this paper seem to contradict that conjecture.

More recent approaches to meeting segmentation have tended to work with richer data representation schemes. Banerjee and Rudnicky [2004] define their model’s data instances as short time windows over meeting segments whose features are described by low-level conversational statistics (number of speakers, number of speaker changes and speech overlap). They train a decision tree classifier to distinguish between windows that contain topic changes, obtaining an 18% accuracy gain over a baseline (random) classifier. In more recent work, implicit supervision in the form of participant notes has been employed in order to segment meetings into speech intervals which correspond to agenda items [Banerjee and Rudnicky 2007]. Dielmann and Renals [2007] segmented meetings from the M4 corpus [McCowan et al. 2003] into a pre-defined set of five basic “group meeting actions”. They used dynamic Bayesian networks to integrate different feature streams (prosody, turn-taking, lexical and video) into a two-level model comprising individual and group actions. Hsueh et al. [2007; 2006] used talk spurts as data instances, assessing the effectiveness of different combinations of features for topic boundary classification, including the above mentioned features as well as prosody and motion data extracted from the video source. They tested feature integration using a C4.5 (decision tree) classifier [Hsueh et al. 2006] and maximum entropy models [Hsueh and Moore 2007]. Although most approaches employ supervised learning, unsupervised learning has also been attempted [Hsueh 2008] using features derived from phonotactic models [Schwarz et al. 2004] or regularities in acoustic patterns [Malioutov et al. 2007] with some degree of success.

The method presented below employs vocalisation matrices as a data representation mechanism for summarising conversational history. An attractive aspect of this approach is that it does not rely on transcribed speech, being therefore unaffected by speech recognition errors. A Naïve Bayes classifier is employed on a combination of continuous and discrete variables [John and Langley 1995], yielding promising segmentation results. The method is described in detail and evaluated in the following section.

4. MDTM SEGMENTATION

There is evidence to suggest that paralinguistic, non-lexical features of speech can be indicative of discourse structure [Grosz and Hirschberg 1992]. Prosodic features, for instance, have been employed as the exclusive means of segmenting speech data from the Switchboard and Broadcast News corpora into sentences and topics [Shriberg et al. 2000]. There is also evidence that the durations of pauses and speech overlaps have predictive value in terms of topic segmentation. Oliveira [2002] notes a correlation between the duration of pauses and topic boundaries in recordings of spontaneous narratives. Statistical analysis of the MDTM corpus shows that boundary and non-boundary vocalisations differ in duration by about 3.9s on average (CI $= [.01, .68]$) (Welch two sample t-test on log transformed values, $t[51.8] = 2.04$, $p < .05$) and that pauses are also significantly longer at topic boundaries (1.5s, CI $= [.02, .38]$, $t[159] = 2.23, p < .05$). Statistically significant differences are also observed in the AMI corpus[3] for pauses (1.2s, CI $= [.68, 1.87]$, $t[597] = 4.19, p < .01$) and vocalisations (7.2s, CI $= [6.4, 8.1]$, $t[919] = 16.7, p < .01$). In terms of the roles performed by the various medical specialists, differences have also been observed. Although medical consultants and pathologists tend to speak at the beginning of PCDs more often than their colleagues (over 44% of boundary vocalisations altogether), their boundary vocalisations are about 4s shorter than their other vocalisations ($p < .01$). On the other hand, medical registrars who also often open PCD with presentations of symptoms and findings spend about 8.3s longer in their PCD boundary vocalisations than in their other interventions CI $= [3.7, 14.2]$, $t[14] = 3.65, p < .01$), in agreement with informal observations reported by Kane and Luz [2009]. These differences suggest that content-free features, combined with participant role information can indeed inform segmentation.

Content-free analysis summarises dialogues as "vocalisation matrices" which basically encode the amount of speech produced by a speaker in a continuous talk spurt, the duration of speech pauses, and the probabilities that a particular speaker's vocalisation will be followed by another speaker's vocalisation. In general, a conversation is modelled as a Markov process with respect to such transition probabilities [Jaffe and Feldstein 1970]. This assumption has been shown to be effective for classification of (pre-segmented) PCDs according to the nature of the discussion (medical, surgical, referral, etc) in [Luz and Kane 2009], where a graph-based representation of the PCD is adopted. The approach adopted here relaxes this assumption by allowing a number of preceding vocalisations to be encoded as part of the feature set.

The data set consists of an interval of silences and vocalisations to be classified as either boundary or non-boundary instances. A boundary instance indicates the beginning of a PCD. The features used to describe an instance are encoded as a vector s encompassing duration of silences or vocalisations and the roles of the speakers who uttered the vocalisations, as shown in equation (1).

[3]More precisely, the sub-corpus corresponding to the "remote control" meeting scenario, with meetings recorded at Edinburgh

$$s = (V_0, L_0, V_{-1}, L_{-1} \ldots, V_{-n}, L_{-n}, V_1, L_1 \ldots, V_n, L_n) \quad (1)$$

V_i is a nominal variable denoting the speaker role (or a pause type or group speech, in the cases of silences and vocalisations by more than one speaker, respectively). The speaker roles which can instantiate $V_0, \ldots, V_n$ range over the values shown in Figure 2(b). Although these roles are specific to MDTMs, other meetings exhibit distinct speaker roles which influence conversational structure [Laskowski et al. 2003]. Recent results suggest that more general roles, such as defined in the AMI corpus, for instance, can be employed for topic segmentation in a similar way as described in this paper [Luz and Su 2010]. Since medical roles denote specialisms it can be assumed that within a stable group like the multidisciplinary team instantiation of the role features can be inferred from speaker identities. Where more than one specialist performed the same role in a PCD (e.g. more than one radiologist took part in the discussion) they were represented as a single role feature as a smoothing technique.

L_i is a continuous variable for the duration of the speech (or silence) interval, and the pairs V_{-i}, L_{-i} and V_i, L_i refer to the i^{th} roles and durations of vocalisation intervals preceding and following the vocalisation described by the instance, respectively.

4.1. Data preparation

As mentioned above, a data set consisting of 54 PCDs has been segmented and manually annotated for speaker identities and roles. In addition to PCDs, the segmentation involved marking the set of dialogue states specified in Definition 4.1.

DEFINITION 4.1. *The following types of* dialogue states *are distinguished:*

. (Individual) Vocalisation*: the length of time that a speaker "has the floor". A speaker takes the floor when they begin speaking to the exclusion of everyone else and speak uninterruptedly without pause for at least 1 second. The vocalisation ends when a silence, another individual vocalisation or a group vocalisation begins. Talk spurts shorter than 1 second (e.g. back channels) are not annotated and are simply incorporated into the main speaker's vocalisation.*

. Group vocalisation *occurs when an individual has fallen silent and two or more individuals are speaking together. The group vocalisation ends when any individual is again speaking alone, or a period of silence begins. Individual speaker identities are lost when a group vocalisation state is entered.*

. Silence *represents quiet periods of over 0.9 seconds between vocalisations (including group vocalisations). A Silence ends when an individual or group vocalisation begins. A Silence can be further classified as:*

  . *a* pause*: a silence between two vocalisations by the same participant,*

  . *a* switching pause*: a silence between two vocalisations by different participants,*

  . *a* group pause*: a silence between two group vocalisations, or*

  . *a* group switching pause*: a silence between a group vocalisation and an individual vocalisation.*

Annotation followed the methodology described in the psychology and computer-supported cooperative work literature [Dabbs and Ruback 1987; Sellen 1995] and therefore focused mainly on amount and structure of speech activity. The metadata created for this set of 54 PCDs are in fact much more detailed, containing information about artifacts employed during the meeting, use of informal language, roles etc. For the purposes of this paper, however, only speech activity and speaker roles are considered. The dialogue states specified in Definition 4.1 are similar to the ones used

by Sellen [1995], with an adjustment to the minimal duration of a vocalisation. Our definition of *silence* is similar to the concept of *switching pauses* [Dabbs and Ruback 1987]. A richer vocalisation event taxonomy could be created through audio sampling at shorter intervals [Brady 1968] but we decided to keep our definition consistent with existing work on group interaction. The threshold of 0.9s in the definition of pause was determined empirically. Sellen [1995], for instance, uses a threshold of 1.5s. However, her data are recorded in 2-participant remote communication scenarios in which pauses tend to be longer due to technology mediation. One could also define simplified notions of *turns* as sequences of vocalisations and pauses, and analogously *group turns* as sequences of group vocalisations and group pauses. However, we chose to avoid the term "turn" altogether, as it is used in conversation analysis [Sacks et al. 1974] in a different and more complex sense.

In keeping the basic units of analysis simple, we expect to be able to automate their extraction from recorded audio through existing signal processing techniques [Fiscus et al. 2008]. It should be noted, however, that the processing steps necessary to turn the audio signal into sequences of dialogue states labelled by speaker (or silence) are not straightforward. This process is called speaker (or audio) diarisation and is usually performed through change detection with the Bayesian information criterion [Chen and Gopalakrishnan 1998] followed by clustering of audio feature vectors using, for instance, Gaussian mixture models as emission probabilities for continuous density hidden Markov models [Ajmera and Wooters 2003]. Although progress has been made in this area [Fiscus et al. 2008; Tranter and Reynolds 2006], diarisation can be quite error prone especially when the input consists of a single audio stream containing all speaker sources. The method described above can therefore be regarded as operating under idealised conditions, in this respect. This simplification seems warranted as a strategy for testing PCD segmentation as an individual module (unit testing). It is also compatible with the approaches to meeting topic segmentation reviewed in Section 3, which for the most part are trained on force-aligned transcripts and speaker identification through individual audio sources (as is the case of the AMI corpus and the ICSI meeting corpus) and therefore approach the standard used in the tests reported in this paper. Nevertheless, it would be interesting to test the method on data containing simulated levels of diarisation error in order to assess its performance under more realistic recording conditions. Asking MDTM participants to wear wireless microphones may also offer a solution which could be tested with the user group in real work contexts.

4.2. Boundary detection

The annotation streams were converted from the ELAN annotation format into an R data frame representing a collection of instances of the form specified in equation (1). Alternative data sets were generated by varying the size of the window over previous and next dialogue state (a horizon of size n role-length pairs on each side of the target dialogue state) and by distinguishing or not between different pause types (see Definition 4.1), in order to assess the effect that these contextual parameters might have on segmentation accuracy.

The segmentation method consisted in training a Naïve Bayes classifier to identify instances marked as boundary dialogue states (i.e. vocalisation that start a new PCD). The conditional probabilities for the nominal variables (speaker roles) are estimated on the training set by maximum likelihood and combined into multinomial models [McCallum and Nigam 1998] while the continuous variables are modelled through Gaussian kernels [John and Langley 1995], as shown in equation (2), where μ_b and σ_b^2 are the mean and variance of the values taken by the features L_i in the data set given

a PCD boundary, represented here as Boolean variable b.

$$P(L_i = x|b) = g(x; \mu_b, \sigma_b) = \frac{1}{\sigma_b\sqrt{2\pi}} e^{-\frac{(x-\mu_b)^2}{2\sigma_b^2}} \quad (2)$$

For the full model, the probabilities to be estimated are simplified through Bayes' rule and the conditional independence assumption to:

$$P(b|S = s) \propto P(V_0 = v_0, \ldots, L_n = l_n|b) \quad (3)$$

$$= \prod_{i=1}^{n} P(V_i = v_i|b)P(L_i = l_i|b) \quad (4)$$

where S denotes a random variable ranging over the vector representation of vocalisation events, as defined in equation (1).

Since the feature sets used in our experiments contained relatively few features, no further pre-processing or feature selection steps were taken during training or classification. The number of PCDs in the test MDTM segments was assumed to be unknown. A maximum a posteriori (MAP) rule [Yang 2001] was adopted for PCD boundary assignment. Other strategies such as SCut and proportional thresholding could also be explored [Luz and Su 2010]. Section 5 discusses thresholding strategies further.

4.3. Evaluation

Although IR metrics such as precision, recall, F scores, and accuracy, have been used to evaluate applications that combine topic segmentation and detection [Banerjee and Rudnicky 2004], the usual way to evaluate meeting segmentation is to employ metrics originally developed for text segmentation. For a segmentation task defined in terms of classification, as in this paper, accuracy scores are misleadingly high due to the fact that the data set is highly imbalanced. Since only about 3% of instances are positive, a trivial classifier assigning non-boundary labels to all instances would predict accurately about 97% of the time. Precision, recall and F scores are also difficult to interpret, even if restricted to the positive class, since they penalise near misses (hypothesised boundaries that fall near true boundaries) and predictions that are wide off the mark equally. Therefore two slightly different error metrics are employed which originated in text segmentation research but are now widely used in speech topic segmentation: P_k [Beeferman et al. 1999] and WindowDiff (WD) [Pevzner and Hearst 2002].

The P_k metric gives the probability that two vocalisations occurring k vocalisations apart and picked otherwise randomly from the data set are incorrectly identified by the algorithm as belonging to the same or to different PCDs. This is formally stated in equation (5), where r and h denote the reference and hypothesis segmentation, respectively. D_k stands for a distribution with probability fixed at a distance k (chosen to be half the average segment size, in number of vocalisations), $a(i, j)$ returns 1 if i and j belong to the same PCD and 0 otherwise, and δ returns 1 if its two arguments are equal and 0 otherwise (Kronecker delta). This results in an increment if boundaries are assigned inconsistently within a segment.

$$P_k(r, h) = \sum_{1 \le i \le j \le N} D_k(i, j)[1 - \delta(a(r_i, r_j), a(h_i, h_j))] \quad (5)$$

The WD metric is based on a similar idea. It can also be regarded as an estimate of inconsistencies between reference and hypothesis, obtained by sliding a window of length equal k segments over the MDTM and counting disagreements. WD, however,

Table II. Mean results for cross-validated segmentation experiments for $\mathbf{1 \leq n \leq 7}$ vocalisation horizons, with and without pause type discrimination. Mean number of boundaries per segment fold is 10 in reference.

	Pause types included			No pause types		
n	P_k	WD	boundaries	P_k	WD	boundaries
1	31.9%	37.0%	6.4	34.6%	39.1%	5.8
2	31.0%	43.0%	14.6	28.4%	44.6%	17.2
3	**27.6%**	38.8%	15.0	28.8%	42.8%	17.2
4	27.8%	35.7%	12.4	27.6%	36.7%	13.6
5	28.1%	**34.7%**	11.2	29.2%	39.3%	13.0
6	31.8%	40.0%	13.2	31.7%	41.6%	14.0
7	31.1%	43.4%	12.6	31.7%	42.8%	13.8

takes into account the number of boundaries predicted by the algorithm and the number actually contained in the reference for the calculation of the error score. The score is calculated as shown in equation (6). N is the number of sub-segments of size k, as before, and $b(i, j)$ gives the number of PCD boundaries between segments i and j.

$$\mathrm{WD}(r, h) = \frac{\sum_{i=1}^{N-k}[1 - \delta(b(r_i, r_{i+k}), b(h_i, h_{i+k}))]}{N - k} \tag{6}$$

In addition to P_k and WD, I follow Sherman and Liu [2008] in reporting the mean number of boundaries actually assigned by the classifier. This is relevant to the interpretation of the results since both segmentation metrics tend to favour hypotheses with fewer boundaries.

4.4. Results

Table II shows the performance of the segmentation algorithm in a 5-fold cross validation experiment in which different window sizes and data representations were compared. Two alternative representations were assessed. In one of them the algorithm distinguished between the various types of pauses specified in Definition 4.1. In the other, it labelled all types of pauses simply as "silence". Results showed that discriminating between pause types (switching pauses, group switching pauses, vocalisation pauses and group pauses) and increasing the vocalisation context horizon both have a positive effect on segmentation accuracy. As the context horizon is increased past 5 vocalisations on each side of the current segment performance degrades as a consequence of data sparsity. Furthermore, it should be remarked that while the single context representation (horizon $n = 1$) results in WD scores close to the best (5-feature horizon) results, its P_k results are clearly inferior. The apparently good performance of the single-feature context in terms of WD is explained by the low average number of PCD boundaries it predicts per segment (see Table II) in conjunction with the fact that the WD metric tends to favour under-prediction. Under-prediction, as will be seen in section 4.7, is one of the main challenges in learning from imbalanced data such as the MDTM corpus. The best performing representations are therefore those which have low P_k and WD values, and yield a number of boundary predictions close to the number in the reference. The 5-feature horizon representation met these criteria better than the alternatives.

A closer analysis of the predictions then reveals that WD scores are considerably higher than P_k scores due to the fact that the algorithm over-predicts boundaries around the true boundary (sometimes predicting as many as 4 hypothetical boundaries adjacently to the true boundary). This is an interesting phenomenon which further supports the hypothesis that the sequential structure of speech exchanges is indicative of higher level (topic) structures. In addition, from a pragmatic perspective, since adjacent boundaries do not occur in practice, this algorithm's behaviour offers a

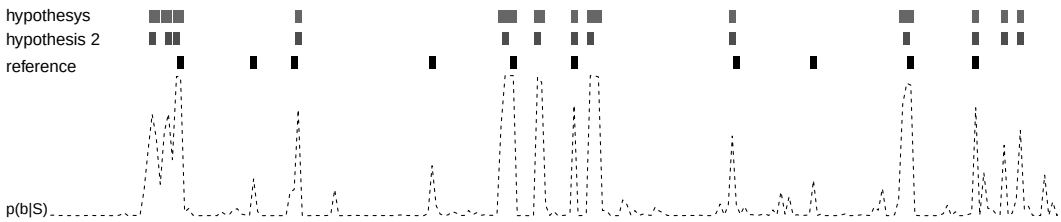

Fig. 3. Profile for a MDTM interval showing segmentation results before (hypothesis) and after (hypothesis 2) removal of spurious adjacent boundaries, in relation to the gold standard (reference). The probabilities assigned to each vocalisation event by the classifier, $P(b|S)$, are shown in the series below the horizontal bars.

straightforward possibility for improvement simply by filtering the excess boundaries at a post-processing step.

Figure 3 shows the segmentation profile for an interval of an MDTM. The filled horizontal bars on the line labelled "reference" represent dialogue events (vocalisations or silences) marked as PCD boundaries in the gold standard annotation. The bars on the line marked "hypothesis" represent the events marked as boundaries by the classifier, possibly containing adjacent event clusters, which could not possibly all be true boundaries. An overview of the probabilities assigned to each dialogue event by the classifier is shown at the bottom of the chart. The line marked "hypothesis 2" shows the boundary assignment results after a simple filtering algorithm was applied which selected among a cluster of adjacent hypothetical boundaries the one with the highest probability (as assigned by the Naïve Bayes classifier) as the true boundary event.

4.5. Baseline analysis

Having observed that PCD (as well as topic) boundaries coincide with vocalisation events that are significantly longer on average than non-boundary events (see page 4), one might wonder whether segmentation based on individual vocalisation events rather than the horizon representation proposed above might not suffice. One would also like to be able to quantify the improvement yielded by the vocalisation horizon representation over reasonable baselines. In this section an analysis of alternative baselines is presented and compared to the results reported above.

Although random and majority classifiers are often used as baselines in machine learning research, they are inappropriate for PCD segmentation due to the imbalanced nature of the data set. Better informed baselines which have been employed in the analysis of transcript-based meeting segmentation analysis include random assignment to the test set of the same number of boundaries found in the training set [Sherman and Liu 2008] and Monte Carlo simulated segments [Hsueh et al. 2006]. Employing a Monte Carlo approach and generating a number of segments proportional to the number in a hold-out MDTM interval, averaged over 100 iterations, gives mean PCD segmentation errors of $P_k = 45.7\%$ and WD $= 50.1\%$. In terms of this baseline, therefore, the optimal results of the horizon technique represent an improvement of about 61.5% for P_k and 69% for WD.

Tests showed that even though vocalisations and pauses tend to be longer at segment boundaries predicting boundaries simply based on vocalisation event duration would not necessarily improve upon the Monte Carlo baseline. Predicting a boundary for all vocalisation events that exceed the mean duration by the amount reported in section 4 would over-predict yielding worse results than the baseline: $P_k = 40.8\%$ and WD $= 51.3\%$. However, a more selective approach, taking only (log-transformed) segments two standard deviations greater than the mean, would produce an improve-

ment: $P_k = 38.5\%$ and WD $= 47\%$. The latter scores show the predictive potential of vocalisation length, but fall well short of the results obtained with vocalisation horizons. Including speaker role information and training a Bayesian model based only on role-duration pairs (i.e. in terms of Table II this would be equivalent to setting $n = 0$ for the horizon) produces similarly unimpressive results: $P_k = 42\%$ and WD $= 46\%$. This analysis shows that considering even a small context (only the immediately preceding and following vocalisation event) considerably improves the predictive power of the representation.

Finally, in attempting to provide a stronger baseline a method was chosen, namely Hidden Markov Models (HMM), which is commonly used for sequence analysis and therefore would appear to be a natural choice for topic segmentation. HMM have in fact been employed for segmentation of telephone and news broadcast speech into sentences [Liu et al. 2006], a task which has some characteristics in common with topic segmentation. A model was created in which b (boundary) and $\neg b$ (non-boundary) corresponded to the model states, speaker roles corresponded to observations, and transition and emission probabilities were estimated from the vocalisation matrix. A 5-fold cross validation procedure was employed for evaluation. The best path hypothesis (Viterbi path) under-predicted yielding $P_k = 38.2\%$ and WD $= 41\%$. In order to mitigate under-prediction, a proportional thresholding strategy [Yang 2001] was applied to the posterior probabilities for b states so as to select a number of boundary instances proportional to the number found in the training set. This strategy resulted in $P_k = 38.7\%$ and WD $= 47.3\%$. These results are further discussed in section 4.7 with respect to the class imbalance issue.

4.6. The effect of diarisation errors

The results above were obtained by training the segmentation algorithm on a gold-standard (i.e. manually annotated and corrected) data set where timing information and speaker identity are reliable across the sequence of vocalisation events. As noted in Section 4, the task of segmenting an audio stream into a vocalisation sequence and assigning these segments speaker labels is known as speaker diarisation. The effectiveness of this task is usually measured in terms of an optimum one-to-one mapping of reference speaker labels to system output speaker labels. A diarisation error rate (DER) metric is then computed as a fraction of speaker time that is mislabelled [NIST 2011].

Although turn-taking boundaries, pauses and overlaps can be reliably identified for dyadic dialogues recorded under favourable conditions (cf. Heldner and Edlund [2010]), diarisation is very much an active area of research. Progress has been made in recent years on diarisation of meeting recordings [Fiscus et al. 2008; Tranter and Reynolds 2006]. The problem, however, is still far from solved, and the speaker diarisation results from the latest rich transcription meeting recognition evaluations [NIST 2011] vary depending on the type of meeting and the audio capture source. Data captured though single distant microphones, for instance, seem harder to process with error rater ranging from 15 to 30%. Multiple distant microphone data, on the other hand, can exhibit diarisation error rates as low as 8%.

Unlike other corpora, the MDTM recordings were taken under challenging acoustic conditions, due to a number of factors. The MDTMs are busy, highly time-constrained events where participants make extensive use of artifacts such as paper records and X-ray films which produce considerable noise. In addition, the video recording equipment (from which one of the audio sources was extracted) had to be placed at the back of the room so as not to interfere with the work of the medical participants. This adversely affected sound quality. In a less exploratory setting where recording would be part of the MDTM routine, multiple microphones could be placed favourably yielding

DER levels comparable to those obtained by current diarisation systems. This section presents an evaluation of the effect of different levels of DER on the PCD segmentation method presented above.

Diarisation consists of various components, including speech detection, change detection, gender and bandwidth classification, clustering, identity finding etc [Tranter and Reynolds 2006]. From the perspective of PCD segmentation according to the method proposed in this paper, the relevant steps are speech and change detection (to determine the duration of vocalisation, pause and overlap events) and clustering/identity finding (to assign speaker labels). As MDTMs have a stable staff membership (the multidisciplinary medical team) with well defined roles, speaker identities map straightforwardly to the role variables ($V_0, \ldots V_n$). In order to access the effects of different levels of DER on segmentation, two types of noise were added to the MDTM data: change detection errors and speaker labelling errors. These noise types were added to reflect the most typical errors found in current diarisation systems. The different errors were assessed both separately and in combination.

Change detection is generally implemented by sliding a window of fixed length over the audio data and looking for changing points within it by using a penalised likelihood ratio criterion (usually the Bayesian information criterion, BIC). A consequence of this approach is that the detector tends to miss short vocalisations (less than 2-5s long) [Tranter and Reynolds 2006]. The distribution of vocalisation lengths in the MDTM data set is highly skewed towards shorter vocalisations (Figure 2a), with about 44% of vocalisations being shorter than 3s. These facts were taken into account when adding noise to vocalisation boundaries so that sorter segments will be more likely to be affected. Thus, vocalisation durations were scaled according to noise drawn from an exponential distribution to target four different levels of DER: 4%, 10%, 17% and 25%, by varying the λ parameter of the distribution. The resulting noisy data sets were tested for actual DER scores using *md-eval* [NIST 2011] and yielded DER values of 4.1%, 9.9%, 17.6% and 26%, respectively.

Speaker error (i.e. where system assigned speaker label differs from reference speaker) accounts to by far the majority of diarisation errors [Tranter and Reynolds 2006; Fiscus et al. 2008]. This type of error was modelled on the MDTM data set by randomly reassigning speaker labels to the vocalisations. The target DER levels were the same as above, and the actual measured scores were: 4.4%, 9.2%, 17.4% and 28%. Finally, we also generated noisy data sets which combined change detection and speaker errors in the proportions reported in recent diarisation evaluations, that is, about 70% of the added noise corresponding to speaker errors and about 30% of noise corresponding to change detection errors. The actual mean DER scores for these data sets were: 5%, 9.2%, 17.5% and 25.8%.

The best performing representation, the 5-vocalisation horizon with pause type discrimination, was chosen as the basis for testing. Data Sets containing diarisation errors in the above defined ranges were converted into that representation and a cross validation procedure was employed over 10 iterations (i.e. noise set generation and segmentation was performed 10 times for each level) and the (P_k and WD) results were averaged. This procedure was repeated for each of the three conditions: *change* detection errors only, *speaker* labelling error only, *combined* change detection and speaker error, as described above.

The results of this evaluation are shown in Figure 4 in terms of P_k and WD scores. As expected, diarisation errors have an adverse effect on PCD segmentation accuracy. However, the technique seems reasonably robust to moderate levels diarisation error as the deterioration in segmentation accuracy is relatively small (6-9.6% in P_k and 14-17% in WD) for DER scores up to 10%. In terms of the effects of different types of errors on segmentation accuracy, change detection had a smaller impact than speaker

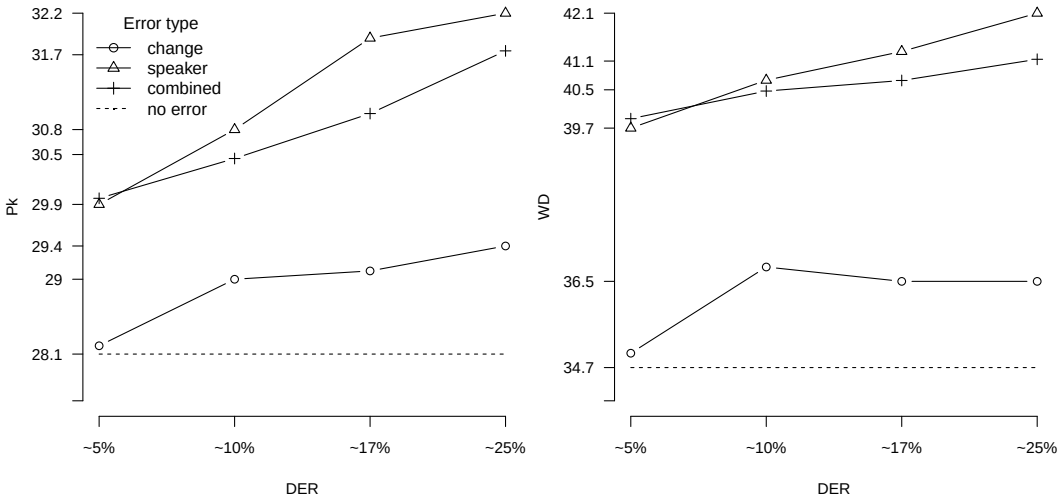

Fig. 4. Segmentation results in P_k (left) and WD (right) for data sets containing diarisation errors. The traced line shows error scores obtained for representation horizon $n = 5$ built on gold standard data (i.e. data containing no diarisation errors).

errors. In the combined error condition, which better reflects errors produced by current systems, segmentation accuracy was similar to that obtained in data sets containing speaker error only. These results imply that the nominal (role label) features of the horizon representation are more useful than their continuous (vocalisation event duration) counterparts.

Despite these encouraging results, a usable system for information access and retrieval MDTM will require accurate diarisation (as well as other "rich transcription" functionality) in addition to robust PCD segmentation. In order to address these needs in a more comprehensive way, we are currently seeking permission to gather a greater volume of medical meeting data using individual microphones and well positioned microphone arrays. These data should enable us to investigate diarisation performance more realistically.

4.7. Redundancy in data representation and the effectiveness of Naïve Bayes

The pattern of boundaries shown in Figure 3, with placement adjacent or clustered around the same regions in the hypothesis (generally around a true boundary) is specific to the use of a Naive Bayes (NB) classifier for the segmentation task. A rerun of the experiment for the best combination of context size 5 and pause type discrimination on three different types of classification algorithms, namely, C4.5 [Quinlan 1993], SVM [Cortes and Vapnik 1995], and nearest neighbour (k-NN) illustrates this point. Table III shows a summary of results in terms of segmentation metrics and boundary numbers. The results for HMM segmentation reported in section 4.5 have also been included for comparison. As can be seen, NB outperforms all other classification methods by a large margin. The difference between the mean number of boundaries initially hypothesised by NB and the mean number of boundaries actually placed after adjacent boundaries were filtered out is particularly noteworthy.

The performance of most classifiers degrades under imbalanced class distributions [Japkowicz and Stephen 2002], as is the case of PCD boundaries in the MDTM data. The class imbalance problem was noted in connection with sentence segmentation and HMM by Liu et al. [2006] who attempted different strategies to mitigate it, including

Table III. Performance of segmentation based on different types of classifiers. Data representation set to a context of 5 vocalisations, including pause type discrimination.

			boundaries	
classifier	P_k	WD	pre-filtering	post-filtering
NB	**28.1%**	**34.7%**	20.4	11.2
k-NN	39.4%	46.3%	7.8	7.6
C4.5	41.0%	46.2%	8.4	8.0
SVM	34.0%	39.0%	5.6	5.2
HMM	38.7%	47.3%	6.0	6.0

a variety of sampling methods in combination with the models. Hsueh et al. [2006] remarked that such strategies appear to be ineffective in meeting topic segmentation where class imbalance is much more severe. In the tests reported here, the decision trees generated by the C4.5 algorithm had to be left unpruned in order for the classifier to avoid trivial classification of all segments as non-boundaries. Similarly the SVM was set to use a simple polynomial kernel, $k(s_i, s_j) = (s_i \cdot s_j)^d$. Despite these adjustments, all tested classifiers with the exception of NB generated on average fewer boundary hypotheses than the reference. As the figures for pre- and post-filtering in Table III show, the clustered placement of boundary hypotheses only occurs with NB.

A possible explanation for the good performance of NB lies in the redundant nature of our data representation scheme. Although the original data representation introduced in Section 4 describes a vector of attributes that correspond to a sequence of vocalisations of a certain length, NB's independence assumption implies that order information is lost when the parameters of the model are estimated. This means that the representations for candidate boundary instances that occur next to each other actually share all but two features.

Zhang [2004] analysed the conditions under which NB can exhibit optimal performance. He concluded that regardless of how strong the dependencies among attributes, good performance can be attained if the dependencies cancel each other out, or are distributed evenly within the classes. In the case of our representation scheme, which always preserves the grouping of roles and vocalisation lengths as it slides a window of fixed size (with respect to the number of discrete vocalisation events, not time) over the dialogue sequence to generate candidate boundary instances, most dependencies will be cancelled out. Furthermore, the similarity among instances in the neighbourhood of a true boundary will have the effect of mitigating the effect of class imbalance. If this is the case, a possibility for improving on the current performance of NB would be to mark non-boundary vocalisations adjacent to true boundaries in the training set as boundaries so as to train the classifier to over-predict around the true boundaries and filter out the excess hypotheses through lower order sequence analysis methods such as HMM. This seems a promising topic for future research.

5. DISCUSSION AND COMPARISONS

The system presented above attains performance levels comparable to those achieved by state of the art supervised systems for segmentation of meetings by topic, while using much simpler content-free features. The decision tree approach presented in [Galley et al. 2003], which is based on lexical cohesion features (LCSeg) extracted from hand-transcribed speech from the ICSI corpus, has error rates of 31.9% (P_k) and 35.9% (WD). The authors report these results to be significantly better than results of other approaches originally designed for text segmentation [Utiyama and Isahara 2001; Choi 2000], whose error scores on the same corpus range from 37.4% to 58%. Sherman and Liu [2008] found that hidden Markov models (over sentence sequences) produces better results than LCSeg on the ICSI corpus, including sub-topics ($P_k = 32.7\%$, WD$= 42\%$).

Hsueh and Moore [2007] report that a lexical cohesion segmentation approach applied to topic segmentation of the AMI corpus produces a P_k score of about 40% and a WD score of 47%. Their improved maximum entropy segmentation algorithm, which combines lexical, conversational, prosody, video and contextual features achieves 34% (P_k) and 36% (WD). These scores were obtained on the task that includes sub-topics, whose ratio of boundary segments to total number of segments is similar to the same ratio observed in the MDTM corpus. The authors also show that moderate levels of word error rates in speech recognition cause only slight degradation in performance, and that not all classes of features are equally important. Somewhat in agreement with the hypothesis investigated in this paper, they find that conversational features are the most essential non-lexical features for topic segmentation.

Although task and corpus differences do not allow a detailed comparison of our results with the ones reported for the above mentioned systems, we note that for a comparable proportion of target boundaries our approach, based solely on amount of speech, speaker transition and role description features, attains lower error rates (27.6% and 34.7% for P_k and WD respectively) than those more elaborate approaches. A similar content-free approach to the one described in this paper, including post-filtering and different threshold strategies as well as further data representation distinctions aimed at better characterising overlaps and pause events, has been tested on the AMI corpus [Luz and Su 2010] yielding relatively good results ($P_k = 27.7\%$ and WD$= 36\%$).

It is likely that PCDs are better structured and homogeneous with respect to turn-taking than topics in more general meetings (even scenario-based ones) and that this structure is captured by our model. In this regard, a comparison to other non-lexical topic segmentation methods which process better structured data such as news broadcasts [Allan et al. 1998] may be helpful. Shriberg et al. [2000] employed decision trees to estimate topic boundary probabilities in broadcast news audio based solely on pause duration, F0 (pitch) range, turn/no turn at boundary, speaker gender and turn duration. They reported error results of about 17.3% in terms of the TDT segmentation metric [Allan et al. 1998]. The TDT metric is an adaptation of the P_k metric [Beeferman et al. 1999] which penalises false alarms more heavily than misses by assigning a 0.7 weight to instances of the former and a 0.3 weight to instances of the latter. This means that in an imbalanced data set the TDT metric is more forgiving than the P_k metric used in this paper (chance in the TDT weighted metric yields a score of 30% while chance on the MDTM yields a P_k of about 45%). In spite of these differences, Shriberg et al's results are impressive and further corroborate the hypothesis that non-lexical features can be good indicators of topic structure. Another hindrance to comparisons between news and MDTM segmentation results is the fact that, even though in both cases the topics have relatively well defined structure, news data are characterised by a more marked contrast between very frequent speakers (e.g. the news anchor) and very rare ones (e.g. interviewees and guests). This kind of contrast is even more evident in data from lectures, to which non-lexical approaches that have also been applied. The unsupervised approach based on audio features only and tested on lecture data by Malioutov et al. [2007] produced a P_k score of 35.8% and a WD of 37%.

From the practical point of view of implementing a searchable multimedia archive of MDTMs [Luz and Kane 2009] usable in a real-world application, segmentation is a very initial but important step. Due to their relatively high error rates, it is unlikely that current segmentation methods could be used for storage of PCD discussion records as separate units on a database system, in a medical context. Rather, we envision an interaction mode in which the user, for instance, "browses" time-based media containing recordings of MDTMs in order to locate the information of interest. The method

presented above, even though it clearly over-predicts, could usefully support this interaction mode. In browsing, high recall is often favoured over precision [Bouamrane and Luz 2007; Moran et al. 1997]. When presented with a misidentified PCD boundary (a false positive), the user can usually identify it as such after a few seconds of listening and skip over to the next boundary. In that regard, it is worth pointing again to Figure 3 and noting that the profile is dominated by zero or very low probabilities (representing true negatives), and that for all missed boundaries (false negatives) the probabilities peak to values greater than those of true negatives. Therefore, if one were to adjust the classification threshold one could optimise the utility of the classifier (in a decision-theoretic sense, valuing recall over precision [Lewis 1995]) for this particular interaction mode. Usability studies to determine and test such parameters are a promising area for future work.

Although the information generated at MDTMs constitutes an invaluable resource for a number of processes in healthcare, from patient management to teaching, the incorporation of MDTM-generated data into existing patient-centred models is far from straightforward. Given that MDT meeting participants work under tight time constraints, automatic recording seems to be the only viable option for data gathering. Recording and storage of multimedia meeting data in digital form have become relatively commonplace in recent years. The challenge consists in finding effective ways of structuring and providing easy access to these data.

6. CONCLUSION AND FUTURE WORK

Collection of meeting data that are representative of the activities of professionals in the real-world is a basic requirement for the analysis of the effects of meeting support technologies on group performance and for the development of systems capable of capturing and indexing of meetings. This paper described a small but, we believe, representative corpus of speech interaction data generated during multidisciplinary medical team meetings[4]. A novel use of a simple data representation technique inspired by research on dialogue in the fields of social psychology and computer supported cooperative work has been presented which produced surprisingly good results in an automatic topic segmentation task. The combination of nominal and continuous features derived from amount and sequence of speech and speaker roles through a Naïve Bayes classifier yielded promising results when applied to the segmentation of MDTMs into patient case discussions, achieving performance levels that compare favourably to state-of-the-art meeting segmentation techniques.

The work described here forms part of an ongoing study aimed at understanding the task and process at play in MDTMs with a view to identifying ways in which computer technology might be deployed in such settings. This includes an investigation into the possibility of enriching existing electronic health records with automatic segmentation and indexing of patient case discussions. Such a system would potentially allow users to easily retrieve PCDs for teaching and healthcare management purposes. In order to achieve these goals, in addition to segmentation, we are currently tackling issues such as automatic categorisation PCDs [Luz and Kane 2009] as well as carrying out further fieldwork studies with the cooperation of the medical teams. In parallel, we are also

[4]Due to the confidential and sensitive nature of the material gathered the audio and video recordings cannot be distributed. However, the anonymised vocalisation sequence data sets and the software used for segmentation and evaluations reported in this paper can be made available on request. In future we hope to obtain approval to gather and distribute privacy-preserving audio features from which content cannot be recovered but that can be used for segmentation at different levels [Parthasarathi et al. 2009] so as to extend the range of possible content-free studies based on the data.

conducting a controlled user study of the usefulness of topic segmentation outputs at different levels of accuracy on a browsing task using AMI corpus data.

As we reach a better understanding of the information needs of the different people involved in MDTM work, we plan on extending the evaluation of our segmentation techniques to encompass the study of empirical correlations between performance at typical information access tasks by, say, senior consultants reviewing similar cases, and the existing segmentation metrics. This could help establish utility criteria based on which parameters such as segmentation thresholds can be tuned. Further work could also explore the detection of specific salient events related to PCD stages. An example of a salient event is the TNM (Tumour, Nodes, Metastases) categorisation by the meeting participants. In addition to MDTM-specific research, we are also carrying out further evaluations on standard meeting corpora.

ACKNOWLEDGMENTS

The author would like to thank Dr Bridget Kane and Dr Su Jing for their collaboration in this study. This study was funded by a Science Foundation Ireland Research Frontiers grant (ECOMMET 06/RFP/CMS054) under the National Development Plan. The feedback and constructive comments of the reviewers and editors is also gratefully acknowledged.

REFERENCES

AJMERA, J. AND WOOTERS, C. 2003. A robust speaker clustering algorithm. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU'03*. IEEE Press, 411–416.

ALLAN, J., CARBONELL, J., DODDINGTON, G., YAMRON, J., YANG, Y., ET AL. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA broadcast news transcription and understanding workshop*. Vol. 1998. 194–218.

BALES, R. F. 1950. *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley, Cambridge, Mass.

BANERJEE, S., ROSE, C., AND RUDNICKY, A. I. 2005. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction (INTERACT'05)*. Springer, 643–656.

BANERJEE, S. AND RUDNICKY, A. 2004. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP)*. ISCA, Jeju Island, Korea, 2189–2192.

BANERJEE, S. AND RUDNICKY, A. I. 2007. Segmenting meetings into agenda items by extracting implicit supervision from human note-taking. In *IUI '07: Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, 151–159.

BARZILAY, R., COLLINS, M., HIRSCHBERG, J., AND WITTAKER, S. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proceedings of the National Conference on Artificial Intelligence*. AAAI Press, 679–684.

BEEFERMAN, D., BERGER, A., AND LAFFERTY, J. 1999. Statistical models for text segmentation. *Machine Learning 34*, 177–210. 10.1023/A:1007506220214.

BLEI, D. M. AND MORENO, P. J. 2001. Topic segmentation with an aspect hidden Markov model. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, New York, NY, USA, 343–348.

BOUAMRANE, M.-M. AND LUZ, S. 2007. Meeting browsing. *Multimedia Systems 12,* 4–5, 439–457.

BRADY, P. 1968. A statistical analysis of on-off patterns in 16 conversations. *The Bell System Technical Journal 47*, 73–91.

BURGER, S., MACLAREN, V., AND YU, H. 2002. The ISL meeting corpus: The impact of meeting type on speech style. In *Seventh International Conference on Spoken Language Processing (ICSLP)*.

CALMAN-HINE, E. 1995. *A policy framework for commissioning cancer services: a report to the chief medical officers of England and Wales. The Calman-Hine Report*. Department of Health, London.

CARLETTA, J. 2007. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation 41,* 2, 181–190.

CHEN, L., ROSE, R., QIAO, Y., KIMBARA, I., PARRILL, F., WELJI, H., HAN, T., TU, J., HUANG, Z., HARPER, M., QUEK, F., XIONG, Y., MCNEILL, D., TUTTLE, R., AND HUANG, T. 2006. VACE multimodal meeting

corpus. In *Proceedings of Machine Learning for Multimodal Interaction (MLMI)*, S. Renals and S. Bengio, Eds. Lecture Notes in Computer Science Series, vol. 3869. Springer Berlin / Heidelberg, 40–51.

CHEN, S. AND GOPALAKRISHNAN, P. 1998. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*.

CHOI, F. Y. Y. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the North American chapter of the Association for Computational Linguistics, NAACL*. 26–33.

CORTES, C. AND VAPNIK, V. 1995. Support-vector networks. *Machine learning 20,* 3, 273–297.

DABBS, J. M. J. AND RUBACK, B. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology 20*, 123–169.

DIELMANN, A. AND RENALS, S. 2007. Automatic meeting segmentation using dynamic Bayesian networks. *IEEE Transactions on Multimedia 9,* 1, 25–36.

FINN, K. E., SELLEN, A. J., AND WILBUR, S. B., Eds. 1997. *Video-Mediated Communication*. Lawrence Erlbaum Associates, Inc.

FISCUS, J. G., AJOT, J., AND GAROFOLO, J. S. 2008. The rich transcription 2007 meeting recognition evaluation. In *Multimodal Technologies for Perception of Humans*. Springer, 3–34.

GALEGHER, J. AND KRAUT, R. E. 1994. Computer-mediated communication for intellectual teamwork: an experience in group writing. *Information Systems Research 5,* 2, 110–139.

GALLEY, M., MCKEOWN, K. R., FOSLER-LUSSIER, E., AND JING, H. 2003. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, E. Hinrichs and D. Roth, Eds. 562–569.

GAROFOLO, J., LAPRUN, C., MICHEL, M., STANFORD, V., AND TABASSI, E. 2004. The NIST meeting room pilot corpus. In *Proc. 4th Intl. Conf. on Language Resources and Evaluation (LREC)*. ELRA, 1411–1414.

GAROFOLO, J. S., AUZANNE, C. G. P., AND VOORHEES, E. M. 1999. The TREC spoken document retrieval track: A success story. In *TREC*.

GROSZ, B. AND HIRSCHBERG, J. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*. ISCA, 429–432.

GROTH, K., FRYKHOLM, O., SEGERSVARD, R., ISAKSSON, B., AND PERMERT, J. 2009. Efficiency in treatment discussions: a field study of time related aspects in multi-disciplinary team meetings. In *Computer-Based Medical Systems, 2009. CBMS 2009. 22nd IEEE International Symposium on*. IEEE, 1–8.

GRUENSTEIN, A., NIEKRASZ, J., AND PURVER, M. 2005. Meeting structure annotation: Data and tools. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*. Lisbon, Portugal, 117–127.

GUTWIN, C. AND GREENBERG, S. 1999. The effects of workspace awareness support on the usability of real-time distributed groupware. *ACM Transactions on Computer-Human Interaction 6,* 3, 243–281.

HACKMAN, J. R. AND MORRIS, C. G. 1975. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In *Advances in Experimental Social Psychology*, L. Berkowitz, Ed. Vol. 8. Academic Press, New York, 45–99.

HEARST, M. A. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics 23,* 1, 33–64.

HELDNER, M. AND EDLUND, J. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics 38,* 4, 555–568.

HINDUS, D., SCHMANDT, C., AND HORNER, C. 1993. Capturing, structuring, and representing ubiquitous audio. *ACM Transactions on Information Systems 11*, 376–400.

HSUEH, P., MOORE, J. D., AND RENALS, S. 2006. Automatic segmentation of multiparty dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. ACL Press, 273–277.

HSUEH, P.-Y. 2008. Audio-based unsupervised segmentation of multiparty dialogue. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 5049–5052.

HSUEH, P.-Y. AND MOORE, J. D. 2007. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL*. Association for Computational Linguistics.

JAFFE, J. AND FELDSTEIN, S. 1970. *Rhythms of dialogue*. Academic Press, New York.

JANIN, A., BARON, D., EDWARDS, J., ELLIS, D., GELBART, D., MORGAN, N., PESKIN, B., PFAU, T., SHRIBERG, E., STOLCKE, A., AND WOOTERS, C. 2003. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*. Vol. 1. 364–367.

JAPKOWICZ, N. AND STEPHEN, S. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis 6,* 5, 429–449.

JOHN, G. H. AND LANGLEY, P. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI'95)*, Besnard, Philippe and S. Hanks, Eds. Morgan Kaufmann Publishers, San Francisco, CA, USA, 338–345.

JORDAN, B. AND HENDERSON, A. 1995. Interaction analysis: Foundations and practice. *The Journal of the Learning Sciences 4,* 1, 39–103.

KANE, B. AND LUZ, S. 2006. Multidisciplinary medical team meetings: An analysis of collaborative working with special attention to timing and teleconferencing. *Computer Supported Cooperative Work (CSCW) 15,* 5, 501–535.

KANE, B. AND LUZ, S. 2009. Achieving diagnosis by consensus. *Computer Supported Cooperative Work (CSCW) 18,* 4, 357–391.

KANE, B. T. 2008. An analysis of multidisciplinary medical team meeting and the use of communication technology. Ph.D. thesis, University of Dublin, Trinity College.

KAZMAN, R., AL-HALIMI, R., HUNT, W., AND MANTEY, M. 1996. Four paradigms for indexing video conferences. *IEEE Multimedia 3,* 1, 63–73.

LASKOWSKI, K., OSTENDORF, M., AND SCHULTZ, T. 2003. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 148–155.

LEWIS, D. D. 1995. Evaluating and optimizing autonomous text classification systems. In *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press, New York, NY, USA, 246–254.

LIU, Y., CHAWLA, N. V., HARPER, M. P., SHRIBERG, E., AND STOLCKE, A. 2006. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language 20,* 4, 468 – 494.

LUZ, S. AND KANE, B. 2009. Classification of patient case discussions through analysis of vocalisation graphs. In *Proceedings of the 11th International Conference on Multimodal Interfaces and Machine Learning for Multimodal Interaction (ICMI-MLMI'09)*. Association for Computing Machinery, ACM, New York, NY, USA, 107–114.

LUZ, S. AND SU, J. 2010. The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In *Proceedings of INTERSPEECH 2010*. ISCA, Chiba, Japan, 1369–1372.

MALIOUTOV, I., PARK, A., BARZILAY, R., AND GLASS, J. 2007. Making sense of sound: Unsupervised topic segmentation over acoustic input. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, 504–511.

MASKEY, S. AND HIRSCHBERG, J. 2006. Summarizing speech without text using hidden markov models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06. Association for Computational Linguistics, Stroudsburg, PA, USA, 89–92.

MCCALLUM, A. AND NIGAM, K. 1998. A comparison of event models for naive Bayes text classification. In *AAAI/ICML-98 Workshop on Learning for Text Categorization*. AAAI Press, 41–48.

MCCOWAN, I., BENGIO, S., GATICA-PEREZ, D., LATHOUD, G., MONAY, F., MOORE, D., WELLNER, P., AND BOURLARD, H. 2003. Modeling human interaction in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*. Vol. 4. 748–51.

MCCOWAN, I., GATICA-PEREZ, D., BENGIO, S., LATHOUD, G., BARNARD, M., AND ZHANG, D. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27,* 3, 305–317.

MCGRATH, J. E. 1991. Time, interaction, and performance (tip). *Small Group Research 22,* 2, 147–174.

MORAN, T. P., PALEN, L., HARRISON, S., CHIU, P., KIMBER, D., MINNEMAN, S., VAN MELLE, W., AND ZELLWEGER, P. 1997. "I'll get that off the audio": A case study of salvaging multimedia meeting records. In *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems*. Vol. 1. 202–209.

MPI. 2005. ELAN: Eucido Linguistic Annotator. Max Planck Institute for Psycholinguistics. http://www.mpi.nl/tools/elan.html.

NIST. 2011. RT evaluation project. http://www.itl.nist.gov/iad/mig/tests/rt/. Retrieved October 2011.

OLIVEIRA, M. 2002. *The role of pause occurrence and pause duration in the signaling of narrative structure*. LNAI Series, vol. 2389. Springer, 43–51.

OLSON, J. S., OLSON, G. M., STORRøSTEN, M., AND CARTER, M. 1993. Groupwork close up: a comparison of the group design process with and without a simple group editor. *ACM Transactions on Information Systems 11,* 4, 321–348.

PARTHASARATHI, S. H. K., MAGIMAI.-DOSS, M., GATICA-PEREZ, D., AND BOURLARD, H. 2009. Speaker change detection with privacy-preserving audio cues. In *ICMI-MLMI '09: Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, New York, NY, USA, 343–346.

PEVZNER, L. AND HEARST, M. A. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics 28*, 19–36.

QUINLAN, J. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

RENALS, S. AND ELLIS, D. 2003. Audio information access from meeting rooms. In *Proceedings of the Internacional Conference on Acoustics, Speech, and Signal Processing*. Vol. IV. IEEE, 744–747.

RENALS, S., HAIN, T., AND BOURLARD, H. 2007. Recognition and interpretation of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*.

RICHTER, H. A., ABOWD, G. D., GEYER, W., FUCHS, L., DAIJAVAD, S., AND POLTROCK, S. E. 2001. Integrating meeting capture within a collaborative team environment. In *Proceedings of the 3rd international conference on Ubiquitous Computing, UbiComp '01*. Springer-Verlag, London, UK, 123–138.

ROBERTSON, T., LI, J., OHARA, K., AND HANSEN, S. 2010. Collaboration within different settings: A study of co-located and distributed multidisciplinary medical team meetings. *Computer Supported Cooperative Work (CSCW) 19*, 483–513.

ROSENBERG, A., SHARIFI, M., AND HIRSCHBERG, J. 2007. Varying input segmentation for story boundary detection in English, Arabic and Mandarin broadcast news. In *Proceedings of Interspeech*. ISCA, Antwerp, 2589–2592.

SACKS, H., SCHEGLOFF, E. A., AND JEFFERSON, G. 1974. A simplest systematics for the organization of turn taking in conversation. *Language 50,* 4, 696–735.

SCHWARZ, P., MATEJKA, P., AND CERNOCKY, J. 2004. Towards lower error rates in phoneme recognition. In *Proceedings of the 7th International Conference on Text, Speech and Dialogue*. 465–472.

SELLEN, A. J. 1995. Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction 10,* 4, 401–444.

SHERMAN, M. AND LIU, Y. 2008. Using hidden Markov models for topic segmentation of meeting transcripts. In *Proceedings of the IEEE Spoken Language Technology Workshop*. 185–188.

SHRIBERG, E., STOLCKE, A., HAKKANI-TÜR, D., AND TÜR, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech communication 32,* 1-2, 127–154.

TRANTER, S. E. AND REYNOLDS, D. A. 2006. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing 14,* 5, 1557–1565.

TUCKER, S. AND WHITTAKER, S. 2005. Accessing multimodal meeting data: Systems, problems and possibilities. In *MLMI '04: Machine Learning for Multimodal Interaction*. Springer-Verlag GmbH, 1–11.

UTIYAMA, M. AND ISAHARA, H. 2001. A statistical model for domain-independent text segmentation. In *ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 499–506.

VINCIARELLI, A. 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia 9,* 6, 1215–1226.

WAIBEL, A., BRETT, M., METZE, F., RIES, K., SCHAAF, T., SCHULTZ, T., SOLTAU, H., YU, H., AND ZECHNER, K. 2001. Advances in automatic meeting record creation and access. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vol. 1. IEEE Press, 597–600.

WHITTAKER, S., LABAN, R., AND TUCKER, S. 2006. Analysing meeting records: An ethnographic study and technological implications. In *Machine Learning for Multimodal Interaction*, S. Renals and S. Bengio, Eds. Lecture Notes in Computer Science Series, vol. 3869. Springer, 101–113.

YANG, Y. 2001. A study on thresholding strategies for text categorization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-01)*, W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, Eds. ACM Press, New York, 137–145.

ZHANG, D., GATICA-PEREZ, D., BENGIO, S., AND MCCOWAN, I. 2006. Modeling individual and group actions in meetings with layered HMMs. *IEEE Transactions on Multimedia 8,* 3, 509–520.

ZHANG, H. 2004. The optimality of Naive Bayes. In *Proceedings of the 7th International Florida Artificial Intelligence Research Society Conference*. AAAI Press.

Received March 2011; revised October 2011; accepted March 2012