An Analysis of Content-free Dialogue Representation, Supervised Classification Methods and Evaluation Metrics for Meeting Topic Segmentation

A Thesis submitted to the
University of Dublin, Trinity College
in fulfillment of the requirements for the degree of

Doctor of Philosophy

Jing Su

April 2011

Declaration

I, the undersigned, declare that this work has not previously been submitted to this or any other University, and that unless otherwise stated, it is entirely my own work. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College.

Jing Su

Date: 29th April 2011

Acknowledgements

I am sincerely grateful to those people and organizations who advised me during the research and rendered generous support.

Dr. Saturnino Luz, my supervisor and mentor, guided me through years of research and significantly advanced my research ability. I benefit from his weekly meeting, in which my understanding and progress were evaluated. He had thorough insight upon the weakness of my work and motivated me along more constructive approaches. Under his supervision, I developed the routine of critical thinking and serious working. This is a lifelong asset on my research career.

Dr. Bridget Kane, my tutor on medical practices, gave me great help to clarify my understanding on hospital work, especially the multidisciplinary team meetings. My work is based on her devoted and pioneering research of MDTMs. It would be not possible to undertake this study without her support and keen advices.

The members and support staff at St. James' hospital warmly supported my research on MDTMs, I sincerely thank all of them. Prof. Donal Hollywood and Dr. Finbarr O'Connell advised me over the need of medical information retrieval, which was important to this study.

My sincere gratitude to Dr. Carl Vogel, Dr. Martin Emms, who reviewed my transfer report and pointed out the weakness of my research as well as the methods to improve it. Dr. Lucy Hederman introduced to me the ECOMMET project and concerned my progress, thanks to her support.

My sincere thank to all persons who had advised me in research, who discussed with me and triggered my ideas. The incomplete list includes Ms. Paula Hicks, Dr. Ielka van der Sluis, Dr. Baoli Li, Dr. Michael Davy, Dr. Mark Matthews, Dr. Jerom Janssen, Dr. Atieh Zarabzadeh, Dr. Dong Zhou, Dr. Xuchun Li, Mr. Ronan McDonnell, Ms. Karen Tierney, Ms. Melanie Späth, Ms. Xiaoge Chen and Mr. Jiang Zhou.

Abstract

Automatic topic segmentation in meeting recordings is intensively investigated due to the fact that topic is a salient discourse structure and it indicates natural reference points for contents. Unlike commonly used text-based topic segmentation methods, this thesis investigates content-free topic segmentation methods. Among the reasons for investigating such methods are: understanding the influence of conversational features in the structure of meeting dialogues, avoiding the complexity of transcription, and protecting confidentiality in sensitive recordings. The research reported here encompasses three major components: classifier selection, sample representation and feature selection, and a set of robust evaluation metrics.

Classification, as a supervised learning method, is employed to distinguish vocalisations that signal topic boundaries from other vocalisations. The unbalanced nature of such vocalisation data sets poses a challenge to commonly used classifiers. However, adapted proportional threshold naïve Bayes classifiers and Boosting classifiers have been found to perform well with proper combinations of vocalisation features. They exhibit segmentation accuracy competitive with text dependent approaches.

Sample representation determines the effectiveness of content-free features. A Vocalisation Event (VE) is proposed as classification unit (instance), in contrast to the fixed length analysis window employed by previous approaches. VE has the advantage of naturally accommodating features such as speaker change, pause, overlap and speaker role. Moreover, VE can be located from audio recordings with speaker segmentation techniques. Experiments show that vocalisation features are more effective than prosody features in topic segmentation.

Based on VE, a Vocalisation Horizon (VH) is proposed as a novel feature concept, in order to indicate temporal or sequence information among classification instances. VE is found to increase segmentation accuracy considerably.

Although P_k and WD are commonly used segmentation metrics, it was found that P_k and WD alone do not suffice to assess the predicted segmentation. A supplemental metric, balance factor ω, is proposed to gauge the ratio of predicted and reference boundaries. Balance factor ω together with P_k and WD support more reliable judgements of segmentation goodness.

These content-free methods were successfully tested on both the Augmented Multiparty Interaction corpus (AMI), which contains simulated meetings, and on the Multidisciplinary Medical Team Meetings (MDTM) corpus, which contains real meetings. MDTMs are better structured meetings than AMI and are segmented with higher accuracy, which indicates the relationship between meeting content and structures.

Contents

1 Introduction **1**

1.1 Background . . . 3

1.1.1 The significant structures of meeting recordings . . . 4

1.1.2 Corpus specific audio structures . . . 5

1.1.3 Content-free topic segmentation method . . . 6

1.2 Thesis Organisation . . . 7

2 Literature Review **9**

2.1 Topic segmentation and topic boundary detection . . . 9

2.2 Text-based topic segmentation . . . 10

2.2.1 Lexical chain and lexical cohesion . . . 10

2.2.2 LCseg algorithm . . . 11

2.3 Text-based topic segmentation with meta-features . . . 13

2.4 Text-independent topic segmentation . . . 17

2.4.1 Pause features . . . 18

2.4.2 Phone and rhyme duration features . . . 20

2.4.3 Pitch features . . . 21

2.5 Speaker segmentation and diarisation . . . 23

2.5.1 Signal Processing . . . 23

2.5.2 Basic Speech Segmentation Techniques . . . 23

2.5.3 Gaussian Mixture Model . . . 25

2.5.3.1 Speaker Segmentation Procedures . . . 27

3 Data set manipulation **28**

3.1 The AMI Corpus . . . 28

3.1.1 Vocalisation event as a unit for topic segmentation 29
3.1.2 AMI vocalisation unit representation 30
3.1.2.1 Empty pauses and overlaps 32
3.1.2.2 Filled pauses 32
3.1.2.3 Acoustic features 33
3.1.2.4 Vocalisation horizon 33
3.2 Multidisciplinary medical team meetings 34
3.2.1 Patient case discussion 35
3.2.2 MDTMs vocalisation unit representation 37
3.2.2.1 Data 39
3.2.2.2 Feature Selection 39

4 System Design 42
4.1 Statistical methods 42
4.1.1 Multiple Logistic Regression 43
4.1.2 Goodness of Fit test 44
4.2 Supervised Learning Algorithms 44
4.2.1 Naïve Bayes 45
4.2.2 Imbalanced data and thresholded Naïve Bayes classifier . 46
4.2.3 Conditional random fields 48
4.2.4 Imbalanced data and ensemble classifiers 50

5 Evaluation Methodologies 52
5.1 Metrics 53
5.2 Near-miss errors 54
5.3 Weakness of popular segmentation metrics 56
5.4 Receiver Operating Characteristics 59
5.4.1 ROC and balance factor 62

6 AMI Experiments 66
6.1 Data set 66
6.1.1 Vocalisation Horizon 67
6.1.2 Features and settings 68
6.2 Classification schemes for topic segmentation 69

6.2.1 Objectives 70
6.2.2 Decision trees 70
6.2.3 Naïve Bayes classifier 71
6.2.4 Thresholding for Naïve Bayes 75
6.2.5 CRF 78
6.2.5.1 Categorised conversational features 78
6.2.5.2 CRF and Naïve Bayes 81
6.2.6 Ensemble classifiers 82
6.2.6.1 The effect of Min instances per leaf in C4.5 . . . 83
6.2.6.2 Using decision tree base classifier 83
6.2.6.3 Using naïve Bayes base classifier 88
6.2.6.4 Conclusion 90
6.3 Feature selection for topic segmentation 91
6.3.1 Empty pause and filled pause 91
6.3.2 Augmented Horizon features 92
6.3.2.1 Speaker role 93
6.3.2.2 Acoustic features 95
6.4 Additional experiments 96
6.4.1 The effect of meeting phases 96
6.5 Conclusion 97

7 MDTMs Experiments 100
7.1 Exploratory Data Analysis 100
7.2 Fitting Statistical Models 105
7.2.1 Model Selection 107
7.2.1.1 Role horizon effects 107
7.2.1.2 VE duration horizon effect 110
7.2.2 Model Validation 112
7.2.3 Segmentation performance 115
7.2.4 Conclusion 118
7.3 PCD segmentation through classification 119
7.3.1 MDTMs vocalisation unit and features 120
7.3.1.1 Using speaker ID or speaker role 121

7.3.1.2 Transforming vocalisation duration 122
7.3.1.3 feature sets 123
7.4 Segmentation experiments 124
7.4.1 naïve Bayes classifier 125
7.4.2 Ensemble classifier 126
7.4.2.1 C4.5 base classifier 126
7.4.2.2 MAP naïve Bayes base classifier 130
7.5 Conclusion 132

8 Speaker Role Identification 135
8.1 Background 135
8.2 Experiments 137
8.2.1 Effect of Different Feature Sets 137
8.2.2 Effect of Vocalisation Horizon 140
8.2.3 Effect of Pauses and Overlap Horizon 141

9 Evaluation of Content-free Topic Segmentation 143
9.1 Migrating from Segmentation to Classification and Regression . . 143
9.2 The advantages of vocalisational events 144
9.3 Effectiveness of segmentation metrics 145
9.4 Measures against skewed and correlated instances 146
9.5 AMI corpus and MDTM corpus 147
9.6 Content-free segmentation and text dependent approaches 149

10 Conclusion 152
10.1 Achievements and innovations 152
10.2 Limitations and future work 154

List of Figures

2.1 F0 processing 21

2.2 Speaker Segmentation Procedures 27

3.1 Schematic Diagram for Vocalisation Horizon, Pause Horizon and Overlap Horizon (Horizon = 3). **Voc** is the current vocalisation, **Vy1** to **Vy3** are vocalisations after **Voc**, **Vz1** to **Vz3** are vocalisations before **Voc**. All 6 instances of vocalisations form the Vocalisation Horizon. In the Pause Layer and Overlap Layer, each instance labels the position of a possible pause or overlap. Between two consecutive vocalisations, there is either a pause or an overlap, or neither. All 6 instances of pauses form the Pause Horizon, and all 6 instances of overlaps form the Overlap Horizon. 34

3.2 Illustration of PCD structure in D-Stage 1 (re-print with permission from Kane [2008]) 36

3.3 From Speaker Segmentation to PCD Segmentation 38

3.4 Topic Segmentation Procedures 38

4.1 Linear-chain CRF 50

5.1 A reference segmentation and five different hypothesised segmentations with different properties. (Adapted from Pevzner and Hearst [2002]) 55

5.2 Confusion matrix 59

5.3 A sample ROC graph with probability thresholds 61

5.4 ROC graph of naive Bayes classifiers on 4 feature sets, MAP-NB and FT-NB correspond to p=0.5 and p=0.99 respectively 64
5.5 *iso-performance line* (slope $m = 99$) and discretised naive Bayes classifiers 65

6.1 C4.5 prediction with EP_{VOCP} features. "reference" indicates the manually assigned topic boundaries and "hypothesis" stands for the predicted boundaries from C4.5 classifier. 71
6.2 MAP Naïve Bayes prediction with EP features. (a) and (b) are from the same data set. In (b), blue dash line indicates posterior probability of predictions in positive class. 73
6.3 Compare MAP and PT Naïve Bayes predictions with EP features, the blue dash line is probability plot (In this figure, instances are different with previous figures, in order to highlight FP predictions, but (a) and (b) are from the same instances and the comparison is valid.) 77
6.4 Density probability distribution of (a)Vocalisation Event duration, (b) Empty pause duration, (c) Filled pause duration and (d) Overlap duration in the unit of Second. Each distribution is skewed and need to be normalised. 79
6.5 Accuracy of ensemble classifiers on EP feature with various C4.5 leaf settings 84
6.6 Compare EP and EP_{VOC} features for C4.5 based AdaBoostM1 (with **NoAdj**) 87
6.7 Effect of **NoAdj** on NB based AdaBoostM1 89

7.1 Vocalisation duration by each speaker 101
7.2 Vocalisation duration by percentile 102
7.3 Log transformed vocalisation duration by each speaker 102
7.4 Log transformed vocalisation duration by percentile 103
7.5 VE duration plot of meeting participants with the same role . . . 104
7.6 Segment boundary plot of multiple logistic regression models . . . 117
7.7 Segmentation accuracy on each fold (5-fold PT NB classification with EP_c features) 121

7.8 Accuracy of ensemble classifiers on EP_c features with various C4.5 leaf settings 127
7.9 Plot of predicted boundaries from winning feature sets in Boosting and Bagging(C4.5 base classifier) 129
7.10 Plot of predicted boundaries from winning feature sets in Boosting and Bagging (naïve Bayes base classifier) 133

8.1 C4.5 classifier accuracy on Speaker Role Classification (left) and Naive Bayes classifier accuracy on Speaker Role Classification (right) 139
8.2 Bayesian Network classifier accuracy on Speaker Role Classification 139
8.3 Speaker Role Classification Accuracy and Vocalization Horizon Effect 140

List of Tables

2.1 Compare the result of MaxEnt models trained with only conversational features (CONV) and with all available features (ALL), (Table 2 from Hsueh and Moore [2007a]) 17

3.1 vocalisation events Feature Set (part of all samples) 40

4.1 Confusion Matrix of MAP NB(a) and C4.5(b) classification results with EP_{VOC} (n=3) feature set from meeting ES2002d. In this table, "a=1" and "b=0" refer to positive and negative instances in reference set respectively; "P" and "N" refer to predicted positive class and negative class respectively. MAP NB predicts 5 true positive instance, but C4.5 predicts no true positive instances. . . 47

5.1 Segmentation accuracy scores P_k, WD and ω in 4 baseline conditions. Test set is a concatenated data set of 30 meetings. The first row in table shows the accuracy scores when no VE is topic boundary in all dataset. The second row shows the accuracy scores when each VE is a topic boundary. The third row stands for the case that the number of predicted topic boundaries is same as the number of true boundaries, and the predicted boundaries are uniformly distributed. The fourth row stands for the same situation as the third row except that the predicted boundaries are randomly distributed. The fifth row stands for the situation that the predicted boundaries have exactly the same quantity and positions as true boundaries. 58

5.2 Segmentation accuracy of MAP and Fixed Threshold Naïve Bayes classifier (Adjacent boundaries are removed) 63

6.1 Segmentation accuracy of C4.5 decision tree on four feature sets with empty pause settings 70

6.2 Segmentation accuracy of MAP Naïve Bayes classifier on four feature sets with empty pause settings (with $NoAdj$) 72

6.3 Classification accuracy of Fixed Threshold and Proportional Threshold Naïve Bayes classifier (with $NoAdj$ filtering). Each classifier is tested on 4 empty pause based feature sets. 75

6.4 Simple statistics on four features of meeting ES2002d: VE duration d, empty pauses duration p_e, filled pause duration p_f, and overlap duration o (unit is Second) 80

6.5 Categorisation of conversational features: (a) d, (b) p_e, (c) p_f and (d) o. In each sub-table, the left column records the interval of a continuous variable, and the right column records the category level assigned to that interval. As a consequence, continuous variables d, p_e, p_f and o are transformed to discrete variables and can be used for CRF training and testing. 80

6.6 Segmentation accuracy from CRF and PT NB 81

6.7 Confusion matrix of four classifiers: (a) CRF, (b) MAP NB, (c) Fixed Threshold NB (p=0.99) and (d) Proportional Threshold NB on EP_{GAP}. (a) CRF has the lowest number of true positive predictions. 82

6.8 Segmentation accuracy of unpruned C4.5 with $M = 15$ 85

6.9 Segmentation accuracy of AdaBoostM1 with C4.5 base classifier, $M = 15$ 85

6.10 Segmentation accuracy of Bagging with C4.5 base classifier, $M = 5$ 85

6.11 Segmentation accuracy of MAP naïve Bayes classifier 88

6.12 Segmentation accuracy of AdaBoostM1 with MAP naïve Bayes base classifier 88

6.13 Segmentation accuracy of Bagging with MAP naïve Bayes base classifier 88

6.14 Classification accuracy of Fixed Threshold (FT) Naïve Bayes classifier on EPs and FPs features (with $NoAdj$) 92
6.15 Classification accuracy of Proportional Threshold (PT) Naïve Bayes classifier on EPs and FPs features (with $NoAdj$) 93
6.16 Segmentation accuracy of PT NB with ROLE Horizon related features 94
6.17 Segmentation accuracy of MAP NB based AdaBoostM1 with ROLE Horizon related features 94
6.18 Segmentation accuracy of PT NB with PITCH Horizon related features 95
6.19 Accuracy of four types of meetings, from EP_{VOC} feature set and with PT NB 97

7.1 Descriptive statistics of continuous variable: **vocalisation duration**101
7.2 Likelihood Ratio Statistics 107
7.3 Multiple Logistic Regression model 1 108
7.4 Chi-square statistics to test the correlation between *Role* and its horizon variables 109
7.5 Multiple Logistic Regression model 2 110
7.6 Pearson correlation test between VE duration $ln(d)$ and its horizon variables 111
7.7 Multiple Logistic Regression model 3 112
7.8 Partition for the Hosmer and Lemeshow test in Model 1 113
7.9 Hosmer and Lemeshow Goodness-of-Fit test on each MLR model 113
7.10 Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c) 114
7.11 Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c), with positive threshold 15% (in training set) 115
7.12 Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c), with positive threshold 15% (in test set) 116

7.13 Confusion Matrix of predictions from logistic regression model 1 (threshould=10%) (a), model 2 (thresold=30%) (b) and model 3 (thresold=20%) (c) (in test set) 116
7.14 Segmentation accuracy of three Logistic Regression models (in training set) 118
7.15 Segmentation accuracy of three Logistic Regression models (in test set with 15% threshold) 118
7.16 Segmentation accuracy of three Logistic Regression models (in test set) 119
7.17 Speaker Roles and corresponding speaker IDs (for the consideration of privacy, speaker ID is replaced by the role code) 122
7.18 Segmentation accuracy of PT-NB with discrete single feature of MDTMs data 122
7.19 Segmentation accuracy of PT-NB with continuous single feature of MDTMs data 123
7.20 Notations of Equation 7.14 to 7.17 124
7.21 PT-NB segmentation accuracy of EP_c and EP_c without st 124
7.22 Segmentation accuracy of PT-NB with MDTMs 125
7.23 Comparison of horizon effect with various naïve Bayes classifiers . 126
7.24 Segmentation accuracy of AdaBoostM1 with C4.5 base classifier ($M = 1$) 126
7.25 Segmentation accuracy of Bagging with C4.5 base classifier ($M = 1$)128
7.26 Segmentation accuracy of C4.5 unpruned decision tree ($M = 1$) . 128
7.27 Segmentation accuracy of MAP naïve Bayes classifier 130
7.28 Segmentation accuracy of AdaBoostM1 with MAP naïve Bayes base classifier 130
7.29 Segmentation accuracy of Bagging with MAP naïve Bayes base classifier 130

8.1 Three feature set combinations 138
8.2 Classification accuracy from three feature sets and three classifiers 140
8.3 Vocalization Horizon effect with VOC_Horizon feature sets on BayesNet and C4.5 classifiers. 141

8.4 Effect of Vocalization Horizon on a BayesNet with GAP_Horizon feature sets. 142

9.1 Statistics on AMI and MDTM corpus 148
9.2 MAP NB classification results in AMI and MDTM corpus 148
9.3 Comparison with other meeting segmentation methods. Only the best reported results are presented. 149

Chapter 1

Introduction

Advances in electronics enable clear voice recording in various working places, where multiparty dialogue recordings contain especially rich information for audience. Although such recordings are valuable for review, people easily get lost in lengthy recordings, since it is hard to locate crucial information without natural reference points. The key is to deliver reference positions or proper index automatically. Research shows that participants took significantly less time to retrieve sought information when they had access to discourse structure based annotation [Banerjee et al., 2005]. Consequently, one distinctly beneficial indexing paradigm is audio discourse structure which minimises the effort in manual searching of favorite content.

Topic is a salient discourse structure indicating context themes. Accordingly it is sought as content guideline by many audience and it is under investigation by many researchers [Galley et al., 2003], [Beeferman et al., 1999], [Hsueh and Moore, 2006]. I consider topic as the most desirable structure for meeting audience to understand the whole recording and locate the parts of interest. In Multidisciplinary Medical Team Meetings (MDTMs) [Kane et al., 2005], Patient Case Discussions (PCD) are regarded as topic and they are the most significant structure of meetings. Each PCD is a well structured event containing diverse aspects of patient diagnosis and management decisions, which are discussed in sequence. It is meaningful to retrieve a PCD in whole.

The AMI corpus is a set of simulated meetings, where teams with predefined roles are engaged in an electrical device design. Each team takes a design from start to prototype in a series of four meetings corresponding to the four phases of the design process Carletta [2007]. With this setting, AMI meetings share common objectives in each phase. Consequently, the objective guides participants along a structured talk and delivers distinguishable topics. These stepwise meetings are desirable for meeting discourse structure retrieval.

Topic segmentation has been intensively studied as an important approach of meeting structure analysis, where most researches are based on meeting transcripts [Galley et al., 2003], [Ries, 2001], [Gruenstein et al., 2005]. Transcription avail topic segmentation with lexical cohesion method [Hearst, 1997], but transcription is not very practical on real meeting (e.g., MDTMs) retrieval, and state-of-the-arts automatic speech recognition is highly challenged with noisy meeting conversations. With regard to this setting, I propose content-free methods for meeting topic segmentation. Acoustic features have been investigated [Hsueh and Moore, 2007b], [Hirschberg and Nakatani, 1998] and are also used in this study. The most inspiring research focuses on conversational features. Psychological researches discover that the amount and structure of vocal exchange influence group interaction [Dabbs and Ruback, 1987], and pausing strategies are used to indicate discourse boundaries [Esposito et al., 2007]. However, vocalisational features have not been utilised independently for automatic topic segmentation purpose. I dedicate in content-free methods and achieve comparable segmentation accuracy as text-based approaches.

In this study, I propose a complete framework of classification schemes based on content-free topic segmentation, and conduct an in-depth study of influential factors. Vocalisation features, including empty pauses and filled pauses are found to be relevant. The definition of a Vocalisation Horizon (*VH*) [Luz, 2009] further improves the segmentation accuracy of an automatic topic segmentation system. Additionally, a study on meeting phases indicates that segmentation accuracy increases with more homogenous meeting contents.

In addition to selecting the right features for topic segmentation, the most challenging task is to find suitable classifiers or adapt classifiers for topic segmentation. The reason is that vocalisation event instances are not independent, but are sampled sequentially. Moreover, the number of instances in boundary and non-boundary classes is highly imbalanced. The thresholded naïve Bayes (*NB*) classifier and ensemble classifiers exhibit advantage over conditional random fields (*CRFs*) and other classifiers in topic segmentation when they are applied to both the AMI corpus and the MDTMs corpus.

1.1 Background

Audio recordings database is a rich source of information, but recordings naturally present linear (or sequential) characteristics which obstruct direct and effective access to the database. The solution emerges from automatically producing structured representation of audio recordings [Hawley, 1993].

In the past decades, numerous dedicated research paved the way of audio processing and understanding. For the domain of speech analysis, we have mature techniques of signal processing (e.g., noise cancellation [Widrow et al., 1975], speech spectrograms [Oppenheim, 1970]), phonetics [Ladefoged, 1993] as well as speech recognition [Junqua and Haton, 1995]. However, most of these techniques focus on information extraction from recordings, and neglect audio structure understanding. For the purpose of producing effective audio content access and retrieval tools, many available speech processing techniques are included (Section 2.4 and Section 2.5).

The major challenge of meeting content access is the linearity of recordings [Luz and Masoodian, 2004; Roy and Luz, 1999]. Many researchers occupy speech recognition techniques and get transcriptions automatically . Based on transcription, linearity of audio recordings is overcome, and various text retrieval methods [Hearst, 1997; Salton and Buckley, 1988] can be applied [Bouamrane and Luz, 2007; Luz and Roy, 1999].

But automatic transcription is not the best choice in many cases. Speech recognition meets with certain level of difficulty in noisy environment, and the generated transcripts actually are absent from intonation and rhythm, which are essential for dialogue understanding. I propose a novel audio structure approximation method: content-free techniques, which are based on acoustic processing outputs instead of transcriptions.

1.1.1 The significant structures of meeting recordings

In order to identify audio structure at topic level, I browse other well organised structure units of text and audio, and explore the relations in between. The features sampled from a fixed level of units (such as vocalisation) can be more rational than that from a drifting window with empirical duration. Furthermore, since topic is a relatively high level structure, it is more feasible to locate topic with lower level structure units than neglecting that. In other words, a stepwise approach in audio topic structure analysis is prefered.

Word entity is a well recognised structure unit of text and speech. Word entity is the foundation of text-based information retrieval, and it is widely applied in transcription based speech analysis. A word entity owns several key properties for audio understanding, that is word meaning, word timing and speed, intonation, etc. All of these properties are of great value for speech analysis, comparing with text mining where only word meaning is considered. Nevertheless, the common method to automatically extract word entities in speech is through speech recognition. Since I prefer text-independent approach for audio understanding, to avoid the errors and computation load of speech recognition, I do not extract word entities from speech. Even if the required features are only word timings and intonation without word meaning, it is necessary to recognise and assemble syllables. As a consequence, word entity is not used as a basic dialogue unit in this study.

Sentence and paragraph are two important discourse unit based on word. They are easily identified in text, but not in dialogue. Pause is a potential indicator of sentence boundary, since it coincides with sentence break in many cases. However, pause is not a reliable sign of sentence boundaries or paragraph boundaries, because it also happens during thinking or other activities before a sentence is finished. As lack of a stable indicator of sentence break and paragraph break, sentence and paragraph are not taken as units in this research of audio structure.

The favorable structure unit in multi-party dialogue recordings is speaker vocalisation turn, which is defined as a piece of continuous talk from a single speaker, with variable length from a few words to several sentences. There are two reasons to emphasise vocalisation turn in audio structure. First, vocalisation turn is highlighting dialogue patterns. The duration of talk, the frequency of turns, and the sequence of speaker talks present rich features to understand a meeting. Second, vocalisation turns can be extracted from recordings with available algorithms and moderate complexity. Speech and non-speech segmentation [Jørgensen and Mølgaard, 2006], speaker change detection [Chen and Gopalakrishnan, 1998] and speaker clustering [Reynolds and Rose, 1995] techniques pave the way to detect vocalisation turns automatically without transcription.

A list of vocalisation turns is not a meaningful index for audience, but it is a building block for topic analysis. Therefore, I explore topic structure based on vocalisation turns instead of word entity, sentence or paragraph. Vocalisation turn is selected as a fundamental unit of recordings, as well as the source of features. Based on vocalisation turns, a variety of text-independent audio accessing methods are investigated, in order to discover the structure of audio content.

1.1.2 Corpus specific audio structures

In previous section, I explore the structure in dialogue/meeting recordings, and text-independent approaches to retrieve it. In that discussion, topics are assumed to exist in multi-party talks. But in reality there can be many examples of structureless dialogues on trivial things. In order to reach a solid conclusion on topic structure, I confine the research in two well-structured meeting corpus: MDTMs and AMI. Common characteristics of these two are clear definition of topics and hierarchical topic settings.

Multidisciplinary Medical Team Meetings (MDTMs) [Kane et al., 2005] have become an established practice in many hospitals, and they have relatively fixed structure. The meeting participants are experts working together, the discussion topics are divided by patient cases, and for each patient, the discussion is organised by certain steps. All of these internal structures are foundation and target of segmentation. The most desirable index item in MDTMs is patient case discussion (PCD). Comparing with word, sentence, paragraph, topic, chapter in articles, a PCD in a meeting is best analogues to a topic in text, so topic boundary detection techniques can be applied to PCD boundary detection.

The AMI corpus Renals et al. [2007] is a collection of simulated meetings with predefined scenario and objectives. So that each AMI meeting is well structured with topics in sequence. Some topics contain subtopic(s). Moreover, different participants are assigned with fixed roles and all audio recordings are transcribed with word-level timings. Speaker turns are helpful feature for segmentation, and it is easy to extract vocalisation boundaries, overlapping and pauses precisely from word-level timings. Based on all of these characteristics, I implement automatic topic segmentation algorithms on the AMI corpus. Comparing with the MDTMs corpus, the AMI corpus is better annotated and open accessible, but they are not real meetings. The segmentation outcome may be influenced by artificial settings.

Since the main objective of this study is to bridge the gap between vocalisation turns and topic structure, I did not implement automatic vocalisation turn detection algorithms. As an alternative, manual annotations of vocalisation boundaries are used to generate vocalisation turns, and the annotation accuracy is satisfying in both corpora.

1.1.3 Content-free topic segmentation method

I aim to achieve automatic topic segmentation for meeting recordings, since topic based index is a favorable meeting structure indicator and it is crucial for meeting review and information retrieval. Different from topic extraction, topic segmentation is mostly achieved through topic boundary detection [Brown and Yule, 1983]. Among the approaches in topic boundary detection, text-based methods exhibit high accuracy and lead mainstream [Galley et al., 2003], [Hearst, 1997]. There are also other successful applications which combine lexical information (e.g., keyword spotting) and conversational features [Hsueh and Moore, 2007b], [Shriberg et al., 2000]. But I am posed to solve topic boundary detection with minimum lexical support, in order to avoid automatic speech recognition errors and reduce computation load.

Without lexical features, the task to find topic structure in multimodal recordings and to offer reference points is challenging. Two schemes are proposed in this study. First, I design topic segmentation algorithms based on other available structures from recordings (e.g., vocalisation turns). One vocalisation turn is treated as a unit with various features, such as speaker identity, duration of talk, and pitch. On the contrary, if fixed length drifting window is used to sample vocalisation features, it is at least difficult to assign speaker identity for a period of talk. Therefore I do not use a fixed length drifting window to sample vocalisation features.

Second, topic segmentation is transformed to a binary classification task. Since it is hard to define one or a set of templates for topic, there is little hope to extract a topic in whole. On the contrary, a supervised learning scheme (binary classification) can be applied to label a vocalisation turn to be a topic boundary or not. The characteristics of a vocalisation turn (e.g., speaker role, talk duration, pauses) are used to train classification models. In classification approach, the primary challenge of topic boundary and non-boundary classification comes from a highly unbalanced data set in two classes, since topic boundary cases only stand for a minor portion of all cases. Furthermore, vocalisation turns are not sampled independently but sequentially, which challenges the assumption of classifiers. In this study I adapt classification schemes to satisfy the skewed data set from audio recordings (Section 4.2). Another research question I faced is that segmentation and classification do not share a common metric for accuracy. A good classification predicts each sample with the right class label, but a good segmentation generates same number of segments as reference, and topic boundaries match reference positions. A set of segmentation metrics are proposed in Chapter 5, in order to accommodate classification approaches.

1.2 Thesis Organisation

There are ten chapters in this thesis including this introduction. In order to give a better understanding of the research field, a review of the topic segmentation related problems is presented in Chapter 2. This review includes a statement of the nature of the problem; a survey of popular text-based topic segmentation methods; a summary of text-independent topic segmentation methods and audio features involved; an exploration of speaker segmentation and diarisation techniques; and a survey of meeting retrieval applications.

In Chapter 3, two meeting corpora for research are presented. The MDTMs corpus contains medical team meeting audio records, and each meeting is composed of a sequence of patient case discussion (PCDs). A PCD is analogues to a general purpose "topic", and PCD boundary detection is performed based on medical staff interactions. Another corpus, the AMI corpus, is a collection of simulated meetings with common scenario and meeting objectives. The differences reside in meeting participants and their style to lead a talk. The AMI corpus benefits topic structure analysis with similar topic sequences and diverse vocalisation patterns. Furthermore, manually annotated hierarchical topics, word level timing and speaker ID labelling in the AMI corpus facilitate research.

Various aspects of experiment design are introduced in Chapter 4. First, statistical models are employed to explore the correlations between topic boundary and other vocalisation features. Then classification methods as well as their adaptations are introduced, to solve data imbalance problem. The AMI corpus and the MDTMs corpus data representations are analysed in the end, and I follow Vocalisation Horizon [Luz, 2009, 2012] as a novel feature to represent relations between samples.

Chapter 5 introduces the metrics for segmentation and defines segmentation fitness. Interesting difference between classification accuracy and segmentation accuracy is noticed, and it leads to a discussion on near-miss errors. Chapter 6 and Chapter 7 present classification experiments in both corpora, and compares the accuracy of classifiers for segmentation purpose. For the same classifier, there is difference of accuracy between the two corpora. In Chapter 9 I discuss the reasons of difference and evaluate the effectiveness of classification schemes. Chapter 10 is the conclusion of this thesis with suggestions for future work.

Chapter 2

Literature Review

2.1 Topic segmentation and topic boundary detection

Topic is informally what is being talked about, and we are accustomed to name the main idea or principle of a talk, or text, as *topic*. However, we do not have a specific criterion to define topic out of a talk or text. As contrast, the notion of *sentence* is easily identified with punctuation. Brown and Yule [1983] stated the difficulty of defining topics in a specific and complete way:

> The notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed to very frequently in the discourse analysis literature ... Yet the basis for the identification of 'topic' is rarely made explicit. (pp. 69-70)

Since it is difficult to identify a topic, they suggest to investigate topic-shift markers as an alternative:

> It has been suggested.., that instead of undertaking the difficult task of attempting to define 'what a topic is', we should concentrate on describing what we recognise as *topic shift*. That is, between two contiguous pieces of discourse which are intuitively considered to have

> two different 'topics', there should be a point at which the shift from one topic to the next is marked. If we can characterize this marking of topic-shift, then we shall have found a structural basis for dividing up stretches of discourse into a series of smaller units, each on a separate topic .... The burden of analysis is consequently transferred to identifying the formal markers of topic-shift in discourse. (pp. 94-95)

Based on the assertion that detecting topic shifts is much more convenient than identifying a topic, numerous approaches are applied on topic shift and people aim to increase prediction accuracy on topic boundaries.

2.2 Text-based topic segmentation

2.2.1 Lexical chain and lexical cohesion

Morris and Hirst [1991] extended the work of Halliday and Hasan [1976], and proposed that *lexical cohesion* is an indicator of discourse structure. Lexical cohesion is explained by chains of related words (lexical chains), which contribute to the continuity of lexical meaning and bridge the gap between vocabulary distribution and discourse structure.

A lexical chain is a manually constructed list of related words in the text, excluding pronouns, prepositions, verbal auxiliaries and high-frequency words. Lexical chains have two main advantages in correlating word entities with text content. First, a lexical chain aids ambiguity resolution and term meaning specification by providing correlated context words. The reason is that a lexical chain is organised mainly from two types of thesaural relations: two words belong to a common thesaurus category, or one word belongs to a sub-category of the other word. So a polysemous word belongs to more than one lexical chain, each owns a different collection of words. In the predefined context span, an algorithm checks the number of correlated words for each hypothesised lexical chain of a polysemous word. A majority vote algorithm determines which lexical chain the polysemous word belongs to and its meaning.

Second, lexical chains provide a clue for the determination of coherence and discourse structure, and hence the larger meaning of the text. Coherence refers to semantic relations within context, such as elaboration, support, cause and exemplification. Generally, it is computationally infeasible to identify coherence relations. Nevertheless, cohesion refers to structural relations on vocabulary level, such as reference, substitution, conjunction and lexical cohesion. Comparatively speaking, cohesion is computationally adaptable. What's more, lexical chains contain words from a common thesaurus category. Since coherence exists in related content, such a collection of words indicates a kind of basic semantic coherence.

Based on these advantages, lexical cohesion is widely applied in discourse structure analysis [Galley et al., 2003], and exhibits convenience in computation.

2.2.2 LCseg algorithm

The TextTiling algorithm devised by Hearst [1997] follows the idea of comparing lexical similarity between the adjacent pairs of text blocks, and computing a *lexical score* between each pair of sentences. Its block algorithm simply computes inner product of two vectors, where a vector contains the number of times each lexical item occurs in its corresponding block. If a low lexical score is preceded by and followed by high lexical scores, this is assumed to indicate a shift in vocabulary corresponding to a subtopic change. In the experiment of locating the subtopic boundaries of 12 texts, TextTiling is shown to produce segmentation that corresponds well to human judgments.

LCseg [Galley et al., 2003] extends TextTiling and performs topic segmentation based on lexical cohesion [Halliday and Hasan, 1976]. LCseg hypothesises that major topic shifts are likely to occur where *strong* term repetitions start and end. Here, a term repetition is a kind of simplified lexical chains (Section 2.2.1), which ignores synonymy and other semantic relations. Strong term repetition indicates high lexical cohesion.

Lexical cohesion score is derived from the similarity between adjacent sampling windows of k words (equation 2.1). At the first stage, it is important to identify and weight strong term repetitions through lexical chains. Equation

(2.3) is a score function combining term frequency and compactness[1], and it is a variant of TF.IDF (term frequency and inverse document frequency) [Salton and Buckley, 1988]. In this equation, $R_1 ... R_n$ is the set of all repeated terms in text, t_i is the corresponding term for R_i, L is the number of sentences in text and L_i is the number of sentences containing t_i. If a term t_i is frequent in a text and it only emerges in very limited sentences, then t_i is a good indicator of lexical cohesion.

Two adjacent windows A and B, each containing k sentences, move along texts with one sentence step each time. Only terms R_i emerging in both windows A and B account for cohesion score (Equation (2.2)). Finally, the cosine similarity is calculated at the transition step to express the score of cohesion, where the cosine similarity is a normalised sum of term scores from windows A and B (Equation (2.1)).

$$cosine(A,B) = \frac{\Sigma_i \omega_{i,A} \cdot \omega_{i,B}}{\sqrt{\Sigma_i \omega_{i,A}^2 \Sigma_i \omega_{i,B}^2}} \tag{2.1}$$

where

$$\omega_{i,\Gamma} = \begin{cases} score(R_i) & \text{if } R_i \text{ overlaps } \Gamma \in A, B \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

and

$$score(R_i) = freq(t_i) \cdot log(\frac{L}{L_i}) \tag{2.3}$$

The lexical cohesion score does not determine topic boundaries directly. The scores are smoothed along sequences and then sentence gaps with low cohesion score are identified as potential topic boundaries.

$$p(m_i) = \frac{1}{2}[LCF(l) + LCF(r) - 2 \cdot LCF(m)] \tag{2.4}$$

where $LCF(x)$ is the value of the lexical cohesion function at x.

A relatively low cohesion score between two adjacent windows may show low similarity in context, but this is not enough to validate the hypothesis that major topic shifts are likely to occur where *strong* term repetitions start and end. Galley

[1] compactness is defined with the length of a term chain. Shorter chains receive a higher compactness score than longer ones, if they contain the same number of terms.

follows Hearst's method [Hearst, 1994] to compare local minima of lexical cohesion score with nearby maxima cohesion score, so as to locate a sentence gap where *strong* cohesion starts and ends. For each local minimum LCF(m_i), the algorithm checks the scores of cohesion to the left of m_i as long as their values are increasing. The maximum value is LCF(l). On the right hand side of m_i, the same algorithm locates maximum as LCF(r). Equation (2.4) determines the probability of a gap being topic boundary, and $p(m_i)$ stands for the sharpness of change in lexical cohesion. The topic segmentation algorithms built on LCseg features (lexical cohesion based) has comparable performance with the state-of-the-art algorithms based on lexical information.

2.3 Text-based topic segmentation with meta-features

Lexical cohesion indicates discourse structure and therefore has been successfully used in text-based topic segmentation. Nevertheless, transcribing meeting recordings and segmenting text is not the only way to achieve topic segmentation. Researchers try to utilise more features beyond meeting transcriptions.

Hsueh et al. [2006] proposed two probabilistic models for topic segmentation, and the study was designed to distinguish the prediction power of both models on top-level topic and subtopic segmentation. The specifications of lexical cohesion based models (LM) and feature-based combined models (CM) models are listed below. Hsueh found that LM achieved competitive results in predicting subtopic boundaries and CM performed best in predicting top-level boundaries, where conversational cues were essential indicators. Since LM features for these experiments were extracted from transcripts, Automatic Speech Recognition (ASR) was tested as an alternative to minimise transcription labor. ASR output had a negative impact on CM models, but did not change the general applicability of the two models.

1. unsupervised lexical cohesion based models (LM)

    - Algorithm: LCSeg

- Feature set: solely lexical cohesion information

2. Feature-based combined models (CM), which is trained on a combination of lexical cohesion and conversation features

    - Algorithm: C4.5 decision tree
    - Feature set:

        (1) lexical cohesion features:

        - the raw lexical cohesion score
        - probability of topic shift indicated by the sharpness of change in lexical cohesion score

        (2) conversational features sampled from a fixed length analysis window covering each potential boundary:

        - the number of cue phrases
        - similarity of speaker activity (the number of words spoken by each speaker)
        - the amount of overlapping speech
        - the amount of silence between speaker turns

In the ICSI meeting corpus [Janin et al., 2003], the top-level topics are defined mostly by meeting structures and general steps, such as "opening", "how to proceed", "closing". On the other hand, the sub-topics are defined by fine-grained meeting contents. For example, "how to proceed" can be subdivided as "data collection", "experimental setup". This two level topic definition is common in meeting copra, including ICSI and AMI. Although the conceptual definitions on topic levels lack precise definition, they have generality across meetings and represents the structure of talk. The difference from Hsueh's LM and CM segmentation output indicated the different probability structure between feature sets and two topic levels. Moreover, the conversational features from CM model were tested to be essential factors on sub-topic segmentation [Hsueh et al., 2006]. Successful application of CM model signals a new topic segmentation approach beyond popular text based segmentation methods. I am interested in the application of conversational features independent of text-based topic segmentation.

The reason is that, the CM model inevitably depends on the LM model to locate potential topic boundaries. Lexical information may increase segmentation accuracy, but constrains the application of text independent features. Hsueh and Moore [2006] extended the topic segmentation task of the ICSI corpus to more general scenario-driven meetings of the AMI corpus. It was showed that CM model outperformed LM model, achieving 20% and 12% improvement on segmenting top-level and sub-topic segments respectively. This result highlights the predictive power of conversational features[2], especially on coarser level topics, although ICSI and AMI segmentation results cannot be compared directly (because evaluation procedures are different).

Hsueh and Moore [2007a] updated the composite topic segmentation approach of Hsueh et al. [2006] and Hsueh and Moore [2006]. The novelty was in four aspects. First, the authors claimed that multiparty dialogue segmentation differs greatly from text segmentation, because multiparty dialogue features group activity and could not be successfully segmented by lexical cohesion. Second, features were not sampled from a fixed length text/speech analysis window but from *spurts*, which were defined as consecutive speech with no pause longer than 0.5 seconds. Third, Maximum Entropy (MaxEnt) classifier was applied instead of C4.5 decision tree, since MaxEnt is a competitive approach in topic and sentence segmentation (Christensen et al. [2005]). The last is an augmented feature set specified below.

1. Five categories of features used for segment boundary recognition (ALL)
    - Conversational Features (CONV)
        - the amount of overlapping speech
        - the amount of silence between speaker turns
        - speaker activity change
        - the number of cue phrases

[2] CM model is a supervised learner with C4.5 decision tree and LM model is an unsupervised learner. Although the difference of segmentation performance may partially explained by supervised learning and unsupervised learning, I attribute the achievement to conversational features. The reason is that conversational features cannot be processed by LCSeg and therefore supervised learning is introduced to topic segmentation.

  - the prediction of LCSEG (i.e., the lexical cohesion statistics, the estimated posterior probability, the predicted class)

- Lexical Features (LX1)
  - the list of words that occur more than once in the spurts that have been marked as a top-level or sub-segment boundary in the training set.
- Prosodic Features (PROS)
  - maximum F0 and energy of spurt
  - mean F0 and energy of spurt
  - pitch contour (i.e., slope)
  - energy at multiple points of a spurt
  - rate of speech (No. of words and No. of syllables per second)
  - in-spurt silence
  - No. of speech pauses preceding and following spurt
  - spurt duration
- Motion Features (MOT)
  - magnitude of relevant movements within a spurt (detected from video in frames of 40ms), which is averaged from close up camera frontal shots, hand movements, presentation areas.
- Contextual Features (CTXT)
  - dialogue act type
  - speaker role

We can see the advantage of *spurts* over fixed length analysis windows in contributing meta features, such as speaker change, silence between speaker turns, preceding and subsequent pauses, speaker role. These features are difficult to extract uniquely and integrally from a fixed length audio window. Table 2.1 shows that the ALL feature set combining 5 feature sets performs better than CONV and LCSEG with each metric, and it is especially useful on top-level topic segmentation. Hsueh further evaluated each feature set in two ways, the first was single feature set experiment, and the second was leave-one-out combination feature sets test. Segmentation results showed that ALL features were significantly better than any single feature set. Furthermore, CONV, PROS, MOTION and CTXT can be taken out from the ALL model individually without increasing the error rate significantly. Finally, LX1+CONV are tested to be most essential feature sets for AMI topic segmentation. They reached $P_k = 0.27$ and $WD = 0.30$ for top-level topics, $P_k = 0.32$ and $WD = 0.35$ for sub-topics. Hsueh's experiments verified the practicality of using conversational features to predict topic boundary, but keywords and lexical cohesion scores were essential in all circumstances. I propose to modify conversational and prosodic features and involve no lexical features in topic segmentation, so as to accommodate content sensitive applications and avoid ASR errors. Moreover, classifier selection and comparison is another essential step we should be aware of. In next section, text independent segmentation methods are introduced.

Table 2.1: Compare the result of MaxEnt models trained with only conversational features (CONV) and with all available features (ALL), (Table 2 from Hsueh and Moore [2007a])

	TOP		SUB	
ErrorRate	P_k	WD	P_k	WD
BASELINE(LCSEG)	0.40	0.49	0.40	0.47
MAXENT(CONV)	0.34	0.34	0.37	0.37
MAXENT(ALL)	0.30	0.33	0.34	0.36

2.4 Text-independent topic segmentation

Although most topic segmentation research is conducted with text-based data sets, text-independent features draw attention and are well studied. In this section, I introduce the work from Shriberg et al. [2000] which contributes prosody-based sentence and topic segmentation methods.

Prosody is a comprehensive concept including speech pause, pitch change, amplitude, melody, and speaking rate variation. Among these specific features, Shriberg choose not to use amplitude- or energy-based features, because they are highly variable among channels and largely redundant with duration and pitch features. Pitch information and durations (empty pause duration and vocalisation duration) are focused on in their study.

2.4.1 Pause features

Pauses are natural components of speech. people usually consider pauses as empty pauses (silent intervals) only, but filled pauses (vocal sounds) are actually another type of pauses. In speech processing, pauses are mostly neglected since they have no lexical meaning. However, pauses contain rich information which are not expressed by transcriptions. For example, a long empty pause may indicate that the speaker is thinking. Moreover, research shows that pauses have communicative functions, such as drawing attention from listeners. Esposito et al. [2007] indicated that pauses are used as a linguistic means for discourse segmentation.

In order to explore topic change with text-independent methods, pauses are important features in this study. The duration of pause is the only property used for pause, and pause duration is extracted at each word boundary. For closely connected words, the pause duration in between is zero.

In this section, I focus on state-of-the-arts pause detection methodologies, which are also named as Voice Activity Detection (VAD).

Zero-Crossing Rate and short-time Energy Jørgensen and Mølgaard [2006] used *Zero-Crossing Rate (ZCR)* to detect the speech sections out of non-speech parts. ZCR is defined as the amplitude sign changes in a frame divided by the length N of the frame:

$$ZCR = \frac{1}{N} \sum_{n=2}^{N} \mid sgn[x(n)] - sgn[x(n-1)] \mid \tag{2.5}$$

Voice signals, as a basic characteristics, have low ZCR value and high Energy; while in unvoiced periods, high frequency random noise is prominent but Energy is low. In this way, ZCR is used in conjunction with short-time Energy to segment speech and non-speech parts. This technique is used by [Lu and Zhang, 2005], [Huang and Hansen, 2004].

Likelihood Ratio Test Sohn et al. [1999] introduced Likelihood Ratio Test (LRT) approach in determining the presence or absence of speech, where the observed signal statistics in the current frame are compared with the estimated noise statistics according to some decision rules.

$$\begin{aligned} H_0 &: \textit{speech absent} : X = N \\ H_1 &: \textit{speech present} : X = N + S \end{aligned}$$

where S, N, and X are L dimensional discrete Fourier transform (DFT) coefficient vectors of speech, noice and noisy speech with their kth elements S_k, N_k and X_k respectively.

Ephraim and Malah [1984] proposed a Gaussian statistical model that DFT coefficients of each process are asymptotically independent Gaussian random variables. Based on this model, the likelihood ratio of the kth frequency band is

$$\Lambda_k = \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} exp\{\frac{\gamma_k \xi_k}{1+\xi_k}\} \tag{2.6}$$

where $\xi_k = \lambda_S(k)/\lambda_N(k)$ and $\gamma_k = |X_k|^2/\lambda_N(k)$ with $\lambda_N(k)$ and $\lambda_S(k)$ denoting the variances of N_k and S_k respectively.

$$log\Lambda = \frac{1}{L}\Sigma_{k=0}^{L-1} log\Lambda_k \begin{cases} < \eta, \ if \ H_0 \\ > \eta, \ if \ H_1 \end{cases} \tag{2.7}$$

The likelihood ratio of current frame equals the geometric mean of the likelihood ratios for the individual frequency bands (Equation 2.7). So that current frame is identified as speech if the total likelihood ratio is higher than threshold η, otherwise noise.

Min/Max Method Esposito et al. [2008] introduced Min/Max algorithm to detect empty pauses. This algorithm is based on the ratio of the predicted minimal noise energy in a detected non-speech region to the maximal noise energy computed on recently detected non-speech regions. The Min/Max ratio allows an adaptation of the generalised threshold level in response to changes in the noise level. The generalised threshold is:

$$T_{k(n)} = \overline{N}_k + [1 - \frac{N_{k,max}}{N_{k,min}(p)}] \cdot \overline{S}_{k,u} \tag{2.8}$$

where T_k is the threshold value computed for kth band and nth frame, $\overline{N}_k$ is the mean of the noise energy computed on recently detected pauses, $\overline{S}_{k,u}$ is the mean value of long-term speech energy computed from the input signal, $N_{k,max}$ and $N_{k,min}(p)$ are, respectively, the maximum noise energy value from recently detected pauses and the minimum noise energy value in the current detected silent pause p.

In addition to the algorithms above, Long-Term Spectral Divergence [Ramirez et al., 2004] and Spectral Flatness Method [Esposito et al., 2008] are also competitive VAD algorithms. Although there are many solutions for empty pause detection, few methods are found on filled pause detection. The latter is a more difficult problem.

2.4.2 Phone and rhyme duration features

Shriberg noted the correlation between sentence/topic boundary and the distribution of phone and rhyme duration. The observations were pre-boundary lengthening, or a slowing down toward the ends of units. Since rhyme is the nucleus and coda of a syllable, the last rhyme preceding the boundary has high potential to be elongated. In order to precisely analyse the extent of rhyme variation, each phone in rhyme is normalised, as in Equation 2.9.

$$\Sigma_i \frac{phone\_dur_i - mean\_phone\_dur_i}{std\_dev\_phone\_dur_i} \tag{2.9}$$

In Equation 2.9, the mean phone duration and standard deviation are calculated from all conversations in the training set. Then the summary of phone duration in a rhyme is divided by the number of phones in it. The mean value of phone duration in the rhyme acts as a prominent feature to notify sentence/topic boundary.

2.4.3 Pitch features

Comparing with pause and phone duration features, pitch incorporates more variation.. Shriberg extracted features from the pitch signal through a four step speech processing algorithm, as shown in Figure 2.1.

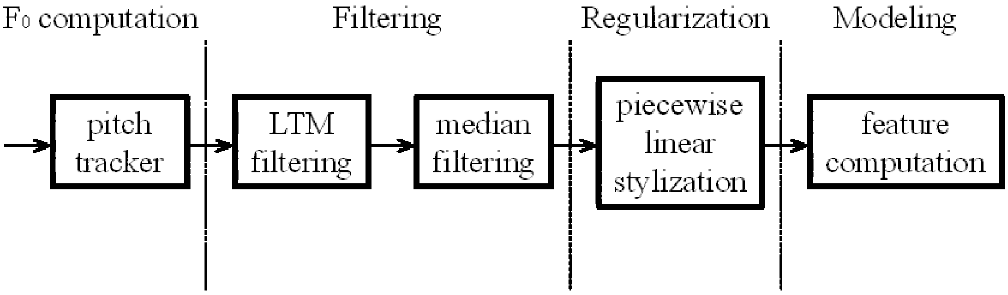


Figure 2.1: F0 processing

The first step, pitch tracker, was a basic smoothing method, which was designed to smooth out micro-intonation and tracking errors, and to simplify F0 feature computation.

The output signal from pitch tracker still contained noise. The "filtering" step further smoothed away noise. A Log-Normal Mixture model (LTM) was trained with each speaker's voice collected previously, and was applied to retrieve each speaker's pitch along recordings (assuming that speaker segmentation has been achieved). Shriberg applied LTM to decompose pitch histogram from each speaker. He used the three log-normal modes placed with log2 space apart to simulate F0 distribution (the centers of three modes are $(\mu - log2, \mu, \mu + log2)$, and the variances in three modes were set to equal). Expectation maximization (EM) algorithm was employed to estimate mixture weights of three modes. With LTM processing I can estimate F0 range parameter of each speaker, for F0 normalization. The last filter in the second step is median filtering. This filter balanced local undershoot or overshoot.

All the previous smoothing and filtering steps were designed to approximate the mainstream of pitch variation. In the third step regularization eventually extracted piecewise pitch slope with a linear model (Equation 2.10). The parameters of linear pitch slope were estimated by minimising the mean squared error.

$$\widetilde{F_0} = \Sigma_{k=1}^{K}(a_k F_0 + b_k) I_{[x_{k-1} < F_0 \leq x_k]} \tag{2.10}$$

The piecewise linear simulation of pitch signals facilitated the last step in F0 processing, which is feature computation. Finally, four features were sampled from pitch. They were F0 *reset* features, F0 *range* features, F0 *slope* features and F0 *continuity* features.

The reset features were identified from the observation that pitch fell at the end of a major unit, and that speakers reset pitch at the start of a new major unit, such as a topic or sentence boundary. So pitch reset could be an important sign of topic change. On the other hand, I should mention that reset features are defined on continuous talk from one speaker. If speaker changes at sentence/topic boundary, reset feature is not valid.

Range features indicated the pitch range of a single word or window from one speaker. Studies showed that features from the word closely before boundary were especially useful for boundary detection. From pitch range I can extract F0 baselines and more, among which F0 baseline is the most useful to identify speaker behavior at boundaries, since speaker voice tends to drop to F0 baseline at topic boundaries.

Slope features precisely represented pitch variations in a word (or window), and they were most useful with continuous vocalisation, such as a word. Continuity features alone were essential indicators for sentence/topic boundary, because sentence/topic shift were unlikely to happen along continuous speech (namely word level, or a short moving window). On the other hand, a discontinuous portion of speech has the potential to be boundary.

The four features extracted from pitch signal are salient indicators for context boundary detection. With the recognised features, Shriberg applied classification approaches, such as decision tree, to detect boundaries. They offered an example of pattern extraction from pitch. Shriberg found that it is not suitable to include original pitch signals as a feature for classification.

2.5 Speaker segmentation and diarisation

As stated in Chapter 1, our preferred text-independent topic segmentation algorithms are based on successful speaker segmentation and clustering. In this section I probe available techniques to automatically achieve speaker segmentation and clustering from raw audio recordings. In following subsections I introduce the techniques related to acoustic processing and post-processing mathematical models.

2.5.1 Signal Processing

The system's input is an audio file, in WAV format (short for Waveform audio format). WAV contains uncompressed audio in pulse-code modulation (PCM) format. To guarantee audio quality, WAV files are sampled at a 44kHz sampling rate and 16 bits per sample.

The waveform of an audio file can represent characters of speech in time-domain. However, time-domain features are much less accurate than frequency-domain features such as Mel-Frequency Cepstral Coefficients (MFCC) (Huang et al. [2001], page 424). Cepstrum is most commonly used frequency-domain feature, but for mel-frequency cepstrum, the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum.

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.11}$$

where $Mel(f)$ is the logarithmic scale of the normal frequency scale f. The Mel scale covers the frequency range of 0 Hz - 20050 Hz.

2.5.2 Basic Speech Segmentation Techniques

Before I can achieve automatic topic segmentation of meeting recordings, there are a few lower level segmentation tasks, which are very important for understanding the basic structure of continuous speech recordings, and are the foundation for higher level speech structure understanding. Fortunately, these tasks have been widely studied during past decades, and the methodologies developed are quite mature.

Speaker segmentation also known as speaker change detection. In continuous speech recording, we need to know if the whole speech is from the same person, and if not, where is the position of speaker change. Chen and Gopalakrishnan [1998] suggests *Bayesian Information Criterion*(BIC) [Schwarz, 1978] as a standard to evaluate the coherence of continuous speech.

The basic idea of BIC comes from maximum likelihood, by which a fixed length speech window is used to sample a cepstral vector along time axis. Assuming each speaker's voice has a distribution of an independent single Gaussian process, I can compare the likelihood that [1,N] samples are from the same speaker (Equation 2.12) or two speakers (Equation 2.13) with change point at time i:

$$H_0 : x_1 ... x_N \sim N(\mu, \sum) \tag{2.12}$$

$$H_1 : x_1 ... x_i \sim N(\mu_1, \sum_1); \quad x_{i+1} ... x_N \sim N(\mu_2, \sum_2) \tag{2.13}$$

The maximum likelihood ratio statistics is

$$R(i) = N \log |\sum| - N_1 \log |\sum_1| - N_2 \log |\sum_2| \tag{2.14}$$

So the maximum likelihood estimate of the changing point is

$$\widehat{t} = \arg\max_i R(i) \tag{2.15}$$

The Maximum likelihood approach has a drawback, it always assign a change point at the highest $R(i)$. Otherwise, I need to predefine an arbitrary threshold to be compared with $R(i)$. To avoid this and make the algorithm robust, BIC integrates the maximum likelihood ratio statistics with a penalty of model complexity. BIC is therefore a thresholding-free method.

Speaker clustering After the speech is segmented and assigned speaker identities, another task is to collect all the pieces of vocalisation that belong to the same speaker; this is a clustering task. Again, BIC can be a good standard to judge the overall fitness of clustering. Chen and Gopalakrishnan [1998] describe one such application of BIC.

Speaker identification Judge a period of unknown voice with the help of identified voice material, a classification task. *Gaussian Mixture Model* (GMM) [Hastie et al., 2009] is the most popular method for this aim. GMM has the advantage to doing the tasks of speaker segmentation and clustering in one step. In next section, GMM is described in detail.

2.5.3 Gaussian Mixture Model

Reynolds and Rose [1995] systematically introduced the methods on speaker recognition, and proposed Gaussian Mixture Models (GMM) [Hastie et al., 2009] for content independent speaker identification. The use of Gaussian Mixture Models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes, in other words, these are acoustic classes. Acoustic classes are useful for modeling speaker identity. The second motivation is, the capability of Gaussian mixtures to model arbitrary densities. This ability is achieved by linear combination of Gaussian basis functions. For acoustic data, a Gaussian mixture density is shown to provide a smooth approximation to the underlying long-term sample distribution obtained from utterances by a given speaker.

A Gaussian mixture density is a weighted sum of M component densities, and given by the equation:

$$p(\overrightarrow{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\overrightarrow{x}) \tag{2.16}$$

where $\overrightarrow{x}$ is a D-dimensional random vector, b_i ($\overrightarrow{x}$), i=1,...,M, are the component densities and p_i, i=1,...,M, are the mixture weights. Each component density is a D-variate Gaussian function of the form:

$$b_i(\overrightarrow{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}}\exp\{-\frac{1}{2}(\overrightarrow{x}-\overrightarrow{\mu}_i)'\sum_i^{-1}(\overrightarrow{x}-\overrightarrow{\mu}_i)\} \tag{2.17}$$

With mean vector $\overrightarrow{\mu}_i$ and covariance matrix $\sum_i$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{p_i, \overrightarrow{\mu}_i, \sum_i\} \quad i = 1, ..., M \tag{2.18}$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ. The spectral shape of the ith acoustic class can in turn be represented by the mean $\overrightarrow{\mu}_i$ of the ith component density, and variations of the average spectral shape can be represented by the covariance matrix $\sum_i$. Because all testing speech is unlabeled, the acoustic classes are "hidden": the class of an observation is unknown. Assuming independent feature vectors, the observation density of observation vectors drawn from these hidden acoustic classes is a Gaussian mixture.

Speaker identification through GMM is performed by supervised learning. In training step, the parameter λ is estimated for one speaker. The most popular training method is maximum likelihood (ML) estimation. The aim of ML estimation is to find the model parameters which maximize the likelihood of the GMM, given the training data. For a sequence of T training vectors $X = \{\overrightarrow{x}_1, ..., \overrightarrow{x}_T\}$, the GMM likelihood is:

$$p(X|\lambda) = \prod_{t=1}^{T} p(\overrightarrow{x}_t|\lambda) \tag{2.19}$$

This is a nonlinear function of the parameters λ. It is not possible to directly maximize the GMM likelihood, but, λ can be obtained by expectation-maximization (EM) algorithm [Dempster et al., 1977].

When the parameters λ are maximized, we can use the identified GMMs from

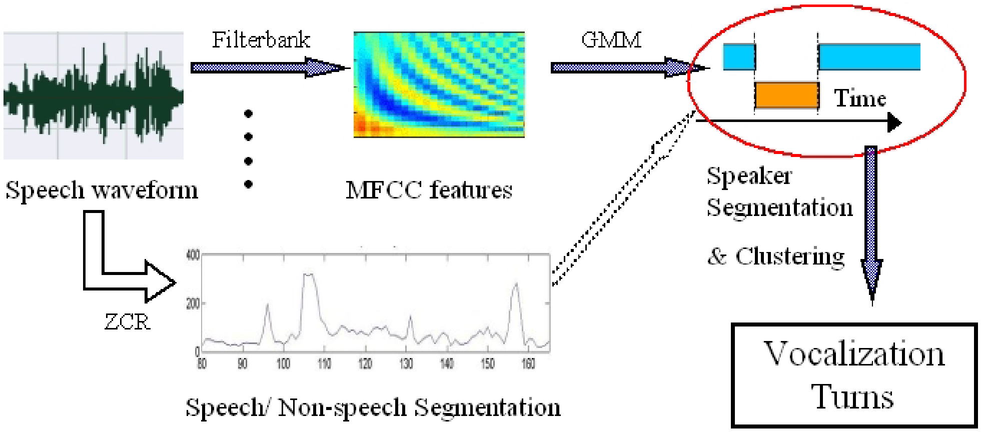


Figure 2.2: Speaker Segmentation Procedures

all speakers to classify a part of unknown speech recording. A group of speakers $\{1, 2, ..., S\}$ is represented by GMMs $\lambda_1, \lambda_2, ..., \lambda_S$. The objective is to find the speaker model which has maximum *posterior* probability for a given observation sequence.

2.5.3.1 Speaker Segmentation Procedures

Figure 2.2 depicts the procedures to get speaker segmentation and clustering. System input is speech recording in WAV format, from which speech and non-speech segmentation is achieved by VAD algorithms; speaker segmentation and clustering could also be achieved by GMM, with MFCC features extracted from WAV. These techniques are introduced in section 2.5.2. At this step, vocalisation turns of each speaker are labeled out of continuous recordings, with the start and end time of each turn. Consequently, content-free topic segmentation models can be constructed with vocalisation turn samples (Section 3.1.1). In this way, topic segmentation models are totally text independent and can be applied on WAV recordings. In experiments, I use vocalisation turn reference from the AMI corpus and the MDTM corpus to examine the accuracy of topic segmentation models.

Chapter 3

Data set manipulation

In this chapter, I introduce two data sets for topic segmentation: the AMI corpus [Carletta, 2007] (Section 3.1) and the MDTM corpus [Kane and Luz, 2006] (Section 3.2). The former is publicly accessible, but the latter is only open to authorised medical staff. Each corpus is introduced of its background, meeting structure, information extraction method, and generated features.

3.1 The AMI Corpus

The AMI corpus is mainly a collection of scenario-based meetings. The motivation for recording such simulated meetings is to define a common framework for a series of recordings and facilitate the research on meeting interactions in a more controlled manner. The fixed key elements are meeting scenario, meeting objectives and meeting roles. The scenario is work meetings of an electronics company to discuss the design of an electrical device. Each participant is assigned to one of the four fixed roles: the project manager (PM), the industrial designer (ID), the user interface designer (UI), and the marketing expert (ME) [Renals et al., 2007]. During meetings, four phases of design procedure are discussed. The objectives of these phases are project introduction, functional design, conceptual design and detailed design. Each meeting is related to only one phase. For these simulated meetings, the common framework helps to control randomness of completely natural languages, and avoid duplication from completely scripted talk [Carletta, 2007].

The AMI corpus offers a convenient resource for language researchers. The corpus is manually transcribed, and annotated with word-level timings. Topic timings are based on word timings. Transcriptions and annotations are recorded in XML format and includes word-level detail for each meeting participant. I parsed meeting annotations to collect features of individual vocalisations. Features include vocalisation duration, vocalisation start and end time. Manually labeled topic and subtopic boundaries are used as reference in classification.

The topics in the AMI corpus are manually annotated after meetings and only have three types: top-level topic, sub-topic, and functional topic. Topics and content of talk are not defined before meetings, but the objective of each meeting is specific. Contents closely related to meeting objective are annotated as top-level topics, inside which clear sub-topics are labeled. Sub-topics in one topic must have common themes, otherwise they would be regarded as independent top-level topics. The AMI annotation scheme only defines one level subtopic. Functional topic refers to transition of meeting (opening or closing part), and can be either top-level or sub-level.

3.1.1 Vocalisation event as a unit for topic segmentation

In order to perform content-free topic segmentation on the AMI corpus, I apply classification and regression methods (Chapter 6 and Chapter 7) with text independent features. Since features are extracted with each instance (or sampling unit), it is essential to generate discrete instances from a continuous recording before modelling. Generally, instances are sampled consecutively or selectively along meeting recording, in a format which is most suitable for the desired features. For example, if speaker ID is a desired feature, an instance must contain only the voice of one speaker.

A drifting window sampler has been applied to extract features for meeting states[1] detection [Banerjee and Rudnicky, 2004], but a window is not capable to represent the properties and patterns from each speaker's voice which contains

[1] Banerjee and Rudnicky [2004] defined two major meeting states as *discussion* and *information flow*. Moreover, *information flow* can be further divided into two states: *presentation* and *briefing*.

essential acoustic and vocalisational features for classification. In other words, a fixed length drifting window does not signify any internal structure of meeting recordings, so as to abandon specific structure information for modelling.

There are many natural structures in meeting recordings, such as word and sentence. But it is difficult to locate these structures without transcription. In text independent experiments, Vocalisation Event(VE) is the term proposed as a classification instance (Section 3.1.2) instead of fixed length moving window. Here a vocalisation event is defined as a period of continuous speech from one speaker, without an empty pause longer than 0.5 second. VE can be extracted with speaker segmentation techniques and only contains voice from one speaker. So it is convenient to extract various vocalisational properties, such as start time, speaker ID, role, duration, pauses and overlaps.

3.1.2 AMI vocalisation unit representation

The AMI corpus contains annotations on the word-level, with timings. In order to resolve overlaps between two speakers and generate proper VE, there are two possible approaches. The first is to terminate a VE when an overlapping vocalisation begins, regardless of whether the current speaker has stopped. A vocalisations generated through this approach is represented as VE_t. The second approach is to continue a vocalisation until the current speaker finishes his/her talk (VE_c). In the AMI corpus, the average duration of VE_t is 1.8s, while the average duration of a VE_c is 4.0s. VE_c is used in this study as a basic unit for classification.

Based on VE_c, I generate eight types of feature sets, and aim to test the effect of feature combination, especially filled pauses (Section 3.1.2.2). If a vocal sound (i.e., “Um”, “Uh”) is regarded as a filled pause, and it is longer than 0.5s, the current vocalisation which contains this vocal sound is split into two new vocalisations, and this vocal sound becomes a filled pause feature p_f for its preceding vocalisation. In this setting, there are three possible observations after a VE_c: empty pause, filled pause and overlap. For simplicity, empty pause, filled pause and overlap are named together as GAP feature g, where $g = (p_f, p_e, o)$.

$$\begin{aligned} FP &= (s, t, d, p_f, p_e, o) & (3.1)\\ FP_{VOC} &= (s, t, d, d_{-n}, ..., d_{-1}, d_1, ..., d_n) & (3.2)\\ FP_{VOCP} &= (s, t, d, p_f, p_e, o, d_{-n}, ..., d_{-1}, d_1, ..., d_n) & (3.3)\\ FP_{GAP} &= (s, t, d, p_f, p_e, o, g_{-n}, ..., g_{-1}, g_1, ..., g_n) & (3.4) \end{aligned}$$

In case Filled Pause is recognised as a proper separator of Vocalisation Event, four Filled Pause (*FPs*) based features are generated as Equations (3.1) to (3.4). Equation (3.1) is a simple feature set containing VE_c speaker s, VE_c start time t, VE_c duration d, VE_c following filled pause duration p_f, empty pause duration p_e and overlap duration o. For one VE_c, there should be only one non-zero value from filled pause or empty pause or overlap following it. If the next VE_c is closely connected, all GAP features are zero.

In Equation (3.3), FP_{VOCP} contains the same features as FP plus Vocalisation Horizon of VE_c duration d. But in Equation (3.2), FP_{VOC} does not contain GAP features p_f, p_e and o. FP_{VOC} is used to signify features of vocalisation itself. In Equation (3.4), FP_{GAP} is analogous to FP_{VOCP} where Vocalisation Horizon is replaced with GAP Horizon. Using these four feature sets with the same classifier, I can easily compare the effect of GAP, GAP Horizon and Vocalization Horizon.

$$\begin{aligned} EP &= (s, t, d_e, p_e, o) & (3.5)\\ EP_{VOC} &= (s, t, d_e, d_{-n}, ..., d_{-1}, d_1, ..., d_n) & (3.6)\\ EP_{VOCP} &= (s, t, d_e, p_e, o, d_{-n}, ..., d_{-1}, d_1, ..., d_n) & (3.7)\\ EP_{GAP} &= (s, t, d_e, p_e, o, u_{-n}, ..., u_{-1}, u_1, ..., u_n) & (3.8) \end{aligned}$$

On the contrary, if a vocal sound is treated as a part of continuous vocalisation, instead of a filled pause, one vocalisation will only stop at an empty pause, and will not be split by vocal sounds.

Equations (3.5) to (3.8) show vocalisation features without filled pauses, where $u = (p_e, o$. I name these features together as EPs based features. In all these equations, s is a unique identifier for a speaker, t is the start time of current VE_c, d is its duration (d_e refers to VE_c duration without filled pause), p_f and p_e are durations of filled pause and empty pause, o is the negative value of overlap duration of adjacent VE_c (in order to distinguish from pause duration), d_i is the duration of the i^{th} VE_c preceding (i<0) or following (i>0) VE_c, g_i is duration of filled pause, empty pause and overlap separately preceding or following VE_c (for VE_c with filled pause), u_i only refers to empty pause and overlap preceding or following VE_c (for VE_c without filled pause) and $n = 3$ is the length of the context (or "horizon") spanned by the VE, as explained in the following section.

3.1.2.1 Empty pauses and overlaps

Pauses can be characterised as *empty pauses* or *filled pauses*. An empty pause corresponds to a period of silence in the conversation. It signals the end of vocalisation or a period of thinking. Beyond these, research shows that pauses have communicative functions, such as drawing attention from listeners. Esposito et al. [2007] indicated that pauses are used as a linguistic means for discourse segmentation. Pauses are used by children and adults to mark the clause and paragraph boundaries. Empty and filled pauses are more likely to coincide with boundaries, realized as a silent interval of varying length, at sentence and paragraph level.

3.1.2.2 Filled pauses

Traditionally, filled pauses are treated as a sign of hesitation and delay. We would like to know how much such hesitations and delays relate to discourse structure. Swerts et al. [1996] analysed acoustic features and shows that filled pauses are more typical in the vicinity of major discourse boundaries. Furthermore, filled pauses at major discourse boundaries are both segmentally and prosodically distinct. Smith and Clark [1993] indicated that dialogue participants have many choices to signal their low confidence on answering questions, and a filled pause is a major option. Speakers use filled pauses to signal that they want to 'hold the floor' [Stenstrom, 1990]. Filled pauses therefore deserve attention, and should be evaluated upon topic boundary detection.

On the selection of filled pause notations, I note that there are two types:

"Um" and "Mm" which have a nasal component, and "Uh" which does not. Clark [1994] showed that in the London-Lund corpus, "Um" and "Mm" are mostly used to signal short interruptions, but "Uh" are used on more serious ones. The two types of filled pauses are analysed separately in this study.

In the AMI corpus, filled pauses are extracted from annotations. "Um", "Mm-hmm", "Uh" and "Uh-huh" are treated as filled pauses exclusively. To be consistent with empty pause extraction, filled pause is identified and extracted only when it is longer than 0.5 second. If a filled pause happens in the middle of a vocalisation event, this vocalisation event is recorded as two vocalisations with a non-switching filled pause in between.

3.1.2.3 Acoustic features

Acoustic features, including *vocalisation speed*, *intensity*, *pitch*, *formant* and MFCC (Mel-frequency cepstral coefficients), are widely used in dialogue and speech analysis. Gaussian mixture models (GMM) achieve reliable speaker segmentation results with MFCC [Reynolds and Rose, 1995], and even LSP (Line Spectrum Pairs), pitch [Lu and Zhang, 2005]. I would like to incorporate acoustic features in content-free topic segmentation. Levow [2004] identified pitch and intensity features that signal segment boundaries in human-computer dialogue, and maximum pitch gathers in segment-initial utterances. I use pitch as an example of acoustic features in topic segmentation research. Praat [Boersma and Weenink, 2009] extracts pitch in the range 75Hz - 600Hz with 10ms sampling rate. I further adapt the pitch recordings with VE_c duration, that is to use the mean pitch value during one VE_c as its pitch feature.

3.1.2.4 Vocalisation horizon

In most classification methods, a vocalisation event is treated as an independent sample. Since the instances are sequentially observed, it is desirable to include time series information into the feature set. I postulate that the features from previous and following vocalisation events can influence the current vocalisation event. I attempt to capture this influence in two ways. The first is by using the duration of previous and following vocalisation events, as a feature of the present

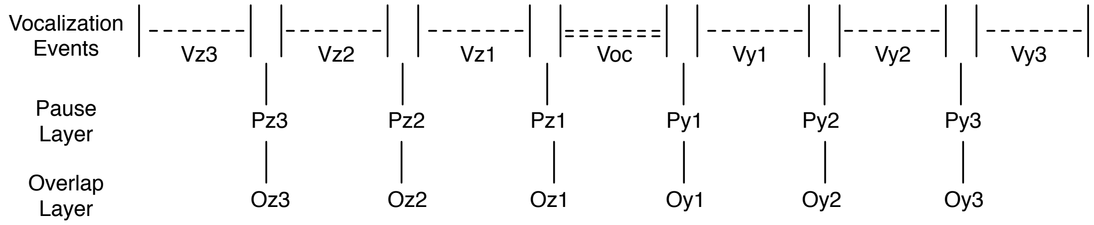


Figure 3.1: Schematic Diagram for Vocalisation Horizon, Pause Horizon and Overlap Horizon (Horizon = 3). **Voc** is the current vocalisation, **Vy1** to **Vy3** are vocalisations after **Voc**, **Vz1** to **Vz3** are vocalisations before **Voc**. All 6 instances of vocalisations form the Vocalisation Horizon. In the Pause Layer and Overlap Layer, each instance labels the position of a possible pause or overlap. Between two consecutive vocalisations, there is either a pause or an overlap, or neither. All 6 instances of pauses form the Pause Horizon, and all 6 instances of overlaps form the Overlap Horizon.

event. We call these features *vocalisation horizon*. The level of vocalisation horizon is the number of vocalisations represented as features on either side of the current vocalisation. For example, level 1 means that only the nearest vocalisation before and after the current vocalisation is used as a vocalisation horizon feature. The second strategy is to use the duration of adjacent pauses[2] and overlaps[3] as features. I call these features *pause horizon* and *overlap horizon* (Figure 3.1). I assume that there is either pause or overlap between any two consecutive vocalisation events. When there is no pause or overlap, the corresponding duration is labeled zero.

3.2 Multidisciplinary medical team meetings

Multidisciplinary Medical Team Meetings (MDTMs) are fora where clinical specialists (that include physicians, surgeons, radiologists, pathologists and oncologists) come together to discuss patient cases. Patient cases are reviewed and decisions are taken on appropriate management for each patient. This multidis-

[2] for *FPs* based feature sets, pauses include empty pause and filled pause, but for *EPs* based feature sets, pauses only include empty pause.

[3] the overlaping vocalisations from two different speakers.

ciplinary method of working in healthcare is advocated by health authorities as it is regarded to improve patient care and reduce the incidence of medical errors over more traditional methods of medical team working [Øvretveit, 2000]. It is anticipated that this system of multidisciplinary team working in healthcare, with meetings as part of that process, will become even more prevalent in routine practice in the future [Kane, 2008].

This method of collaborative working through speech interaction poses challenges for organisation and patient record keeping. Traditional models of the electrical patient record (EPR) have failed to take account of the reality of the setting where the work actually takes place [Hartswood et al., 2003]. A drive is evident towards integrated care records for patients that take account of the collaborative dimension of health care work and support informality in ways that are organisationally acceptable [Hardstone et al., 2004]. However, MDTMs are organised with a robust internal structure, together with rhythms of execution of pre-meeting and post-meeting activities. Even if the communication and meeting coordination are retarded in teleconferencing, MDTMs structure keeps stable [Kane and Luz, 2006]. The stable meeting structure is a salient characteristic which facilitates the study of MDTMs recording retrieval.

3.2.1 Patient case discussion

MDTMs can be described as multiparty meetings and the most important subunit in these meetings is an individual Patient Case Discussion (PCD). PCD is a concept on topic level. Although PCDs have a lot common characters (such as discussion stages), each individual PCD is different in content, length, etc. For the purpose of reviewing an MDTM recording or for medical education, materials segmented by PCD will be most favorable. To satisfy this need, the first aim of automatic MDTMs segmentation is set to segment by PCD, and finally a robust audio index is built on PCD level. With such index, audio retrieval/ review will be greatly facilitated. From the hierarchical structure of audio file, the most effective solution should come from the adjacent level, vocalisation events. If one vocalisation event is regarded as the first talk of a PCD, the boundary of PCD is determined, and the segmentation task is achieved. The problem to distinguish

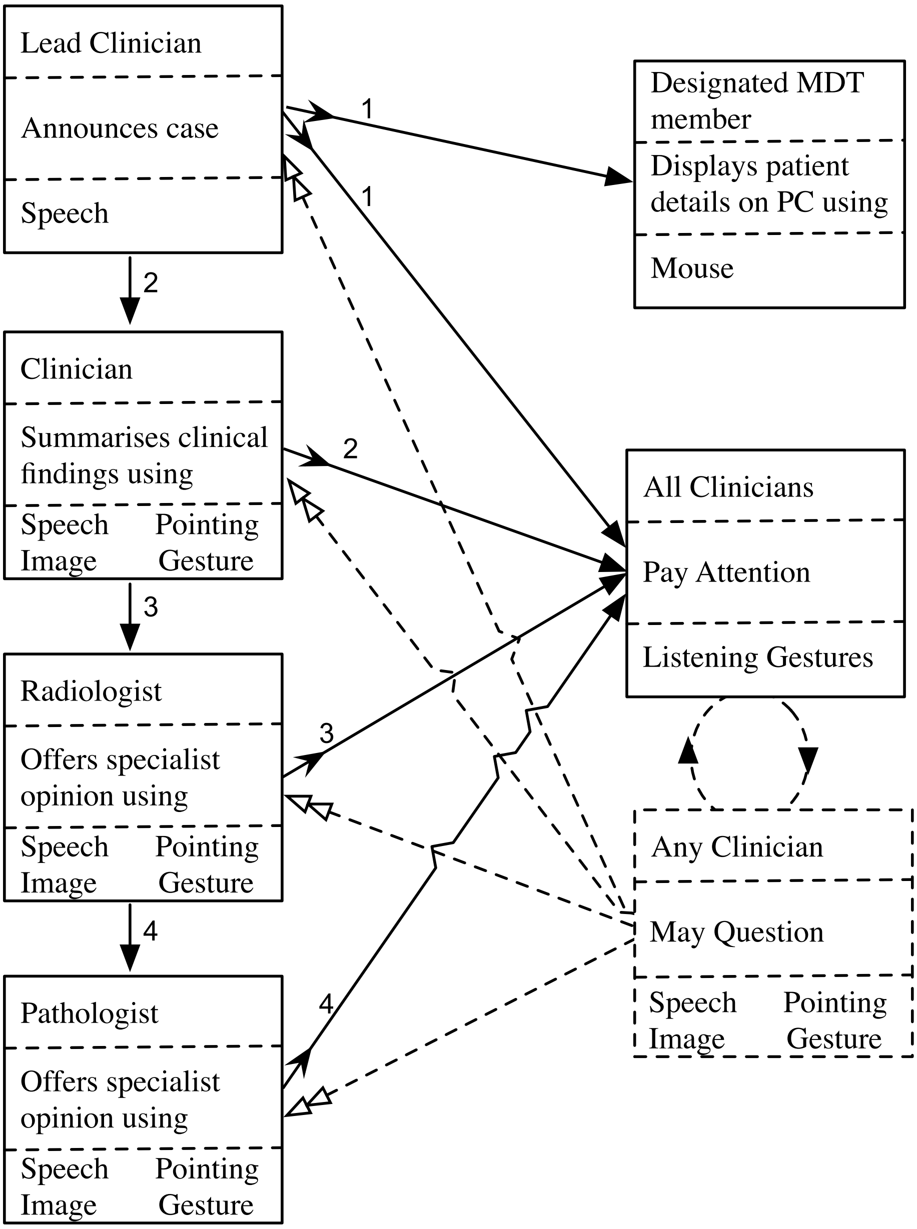

Lead Clinician
Announces case
Speech
Designated MDT member
Displays patient details on PC using
Mouse
1
1
2
Clinician
Summarises clinical findings using
Speech Pointing Image Gesture
2
All Clinicians
Pay Attention
Listening Gestures
3
Radiologist
Offers specialist opinion using
Speech Pointing Image Gesture
3
Any Clinician
May Question
Speech Pointing Image Gesture
4
Pathologist
Offers specialist opinion using
Speech Pointing Image Gesture
4

Figure 3.2: Illustration of PCD structure in D-Stage 1 (re-print with permission from Kane [2008])

the vocalisation events starting a PCD from those not at the start, is a binary classification problem. Figure 3.3 shows the relation between vocalisation events and PCDs.

Individual PCDs vary considerably in length. The duration of a PCD is shown to vary by over 50% of the mean PCD duration, depending on the nature of the patient's clinical presentation [Kane, 2008]. The variety of PCD lengths is a big challenge to segmentation algorithms. However, although the duration can vary, the interaction structure among the specialist roles is relatively stable and predictable, and thus is amenable to automatic generalization. For example, clinicians usually introduce a patient's situation in early part of a PCD, his/her talk will be important to trace the start time of a PCD (as indicated in Figure 3.2). But, clinicians may also talk a lot in the middle of a PCD. In order to handle this complex situation, many features will be involved in PCD segmentation, and various data mining techniques will be applied consequently. Moreover, since specialist roles and vocalisation events are potential indicators of PCD structure, I prefer vocalisation events over a fixed length sampling window for topic boundary detection.

A further aim of topic segmentation in MDTMs is to automatically analyze the sub-topic structures in one PCD. A PCD is well structured and can be divided into four stages [Kane, 2008]. These stages may interest some specialists, and the research on stages may enhance PCD segmentation.

3.2.2 MDTMs vocalisation unit representation

Figure 3.3 shows our objective to overcome the barrier from *speaker segmentation* to *topic segmentation* on MDTM recordings, in an automatic way and without transcription.

In order to guarantee the best result of topic segmentation, I employ manual speaker segmentation results (vocalisation events) as input to topic segmentation models. In the following sections, data analysis and segmentation strategies will be proposed. Figure 3.4 shows feasible procedures from speaker segmentation to content-free topic segmentation.

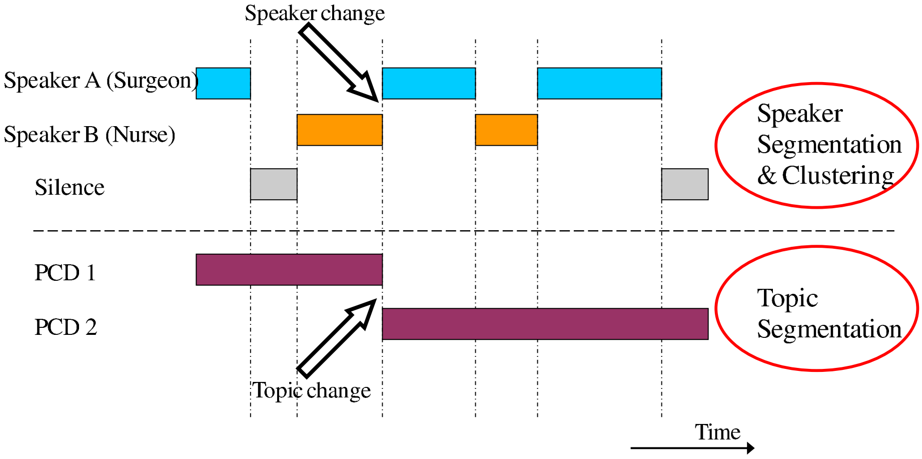


Figure 3.3: From Speaker Segmentation to PCD Segmentation

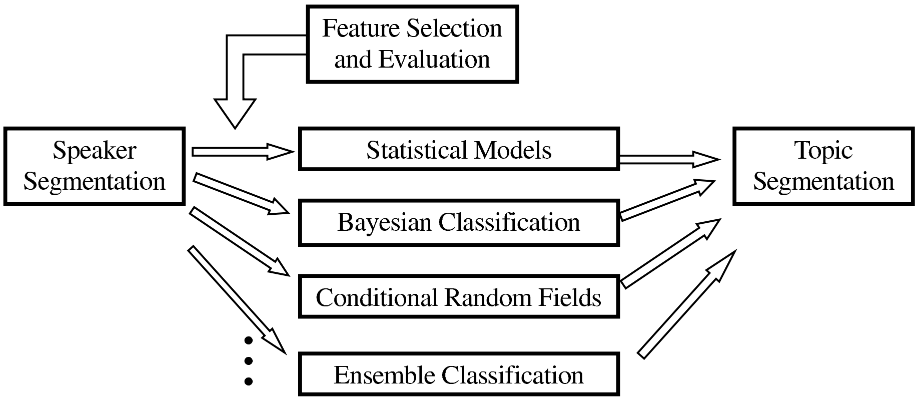


Figure 3.4: Topic Segmentation Procedures

3.2.2.1 Data

In order to process the MDTM meeting recordings, the first step is to extract acoustic features from speech. All segmentation strategies in this study are based on acoustic features. In this study, the hidden Markov Model Toolkit (HTK) [Young, 2007] is used to extract Mel Frequency Cepstral Coefficients (MFCC). MFCC has an advantage over Fast Fourier Transform (FFT) in approximating human auditory system's response, because the frequency bands are placed logarithmically in Mel-frequency cepstrum. Audio recordings are sampled in 10msec rate and MFCC is extracted from each Hamming window of 25msec duration. The most significant 12 dimensions of cepstral coefficients are taken as vectors of acoustic features in recordings. In this way, audio recordings are transformed into feature vectors. These features represent the distinguishable acoustic characters of meeting recordings on frequency domain, and are used as the basis for speaker segmentation.

The second step is speaker segmentation, i.e. to segment the speech into pieces. Each piece is either voice from one speaker or silence, or noise.

As pointed out in Section 2.5.2, speaker identification technique has advantage in achieving speaker segmentation and speaker clustering in one step, when Gaussian Mixture Models (GMMs) are utilised [Reynolds and Rose, 1995]. The method consists of manually collecting voice samples from each speaker, and using them to train a GMM model for that speaker. Then GMM models for all speakers are applied to the meeting recordings. Repeatedly select 10msec data along time sequence, calculate likelihood to each GMM, and assign data to the model (speaker) with the highest likelihood.

3.2.2.2 Feature Selection

As stated in Section 3.1.1, vocalisation events are regarded as the basic unit for classification and segmentation models in the AMI corpus. For MDTM topic segmentation, VE is also used as a sampling unit instead of other standards. Moreover, the PCD segmentation task is transformed to PCD boundary detection task. I use a binary variable $CaseB$ to indicate a PCD start. If one VE is the beginning of a new case discussion, $CaseB = 1$, otherwise $CaseB = 0$.

The dataset of vocalisation events is a time series with the Beginning and End time of each vocalisation slot, and speaker name of corresponding slot. Samples from this dataset are shown in Table 3.1. I use it as training set, so it contains *caseno*, which is label of each case discussion. Our objective is to separate PCDs automatically, with all features available. PCD separation is achieved via a binary variable *caseB*, indicating a PCD start vocalisation and other vocalisations. In a Multiple Regression Model, *caseB* is a standard response. A formal representation of the MDTM classification feature sets is in Section 7.3.1.3.

Begin	Duration	Speaker	Role	Caseno	caseB	before1	next1
0	45.83	V	SURG	1	**1**	.	B
45.83	1.76	B	PHYS	1	**0**	V	V
47.59	36.422	V	SURG	1	**0**	B	J
84.012	1.72	J	RADI	1	**0**	V	V
85.732	16.26	V	SURG	1	**0**	J	B
101.992	7.988	B	PHYS	1	**0**	V	V
109.98	5.018	V	SURG	1	**0**	B	B
114.998	6	B	PHYS	2	**1**	V	V
120.998	1.99	V	SURG	2	**0**	B	JP
122.988	13.49	JP	PATH	2	**0**	V	G

Table 3.1: vocalisation events Feature Set (part of all samples)

The proposed variables in the Multiple Regression Model are:

CaseB: binary response, indicates the start of each case discussion

Duration: time length for each vocalisation slot

Speaker: name of the speaker matching each vocalisation event (VE)

Role: categorical variable of each speaker's role, including Surgeon, Physician, Radiologist, Pathologist, Oncologist, Nurse

Before1: the name (can be replaced by role) of the previous speaker

Before2: the name (can be replaced by role) of the second previous speaker[4]

[4] In this sample feature set, Vocalisation Horizon=2, so the third previous speaker and the third next speaker are not included.

Next1: the name (can be replaced by role) of the next speaker

Next2: the name (can be replaced by role) of the second next speaker

Chapter 4

System Design

This chapter explains the system design and methodologies of content-free topic segmentation. I propose multiple logistic regression model [Agresti, 2002] to explore the relations between vocalisation features and topic boundary at the first stage. Section 4.1 introduces the regression model and the method of Goodness-of-Fit test. Section 4.2 systematically introduces several classification methods, including naïve Bayes, conditional random fields, and ensemble classifiers which have been applied to segmentation. Before presenting the segmentation experiments in following chapters, I present single features and feature sets of the AMI corpus and the MDTM corpus respectively in Section 3.1.2 and Section 3.2.2.

4.1 Statistical methods

In this section I introduce statistical models for exploratory data analysis in the MDTMs scenario. A prominent feature of MDTMs is that there are many meeting participants, varying from 10 to 20. People with the same specialities perform the similar duties, so it is interesting to find the relation between speakers' roles at the meeting and the possibility of a topic-start vocalisation. Other features such as vocalisation duration are also of interest.

A Multiple Logistic Regression model (MLR) expresses the linear relationship between independent variables (numerical or categorical) and a binary response variable (probability of a binary choice) (Section 4.1.1). Goodness of fit test evaluates the fitness of a regression model (Section 4.1.2).

4.1.1 Multiple Logistic Regression

Suppose that I have n binomial observations of the form y_i/n_i, for i=1,2,...,n, where the expected value of the random variable associated with the ith observation y_i, is given by $E(Y_i) = n_i p_i$ and p_i is the corresponding response probability. The linear logistic model for the dependence of p_i on the values $x_{1i}, ..., x_{ki}$ of k explanatory variables, $X_1, ..., X_k$, is:

$$logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} \tag{4.1}$$

which can be transformed to:

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})} \tag{4.2}$$

In the formula 4.1, β_i is estimated by maximum likelihood. For MDTM segmentation, not all of the features of a vocalisation turn are influential factors. The steps to select influential factors in formula 4.1 are called model selection. Model selection is performed in decreasing order. At first the saturated model is fitted, then redundant variables are removed step by step, until all variables are significant. Then the factors in the final model are selected.

Since y_i is an observation on a binomial random variable Y_i, with mean $n_i p_i$, a corresponding model for that expected value of Y_i is:

$$\begin{aligned} E(Y_i) &= n_i \exp(\eta_i)/[1 + \exp(\eta_i)] \\ \eta_i &= \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} \end{aligned} \tag{4.3}$$

The *linear logistic model* is a member of a class of models known as *generalized linear models*, introduced by Nelder and Wedderburn [1972]. This class also includes the linear model for normally distributed data and the log-linear model for data in the form of counts. The function that relates p_i to the linear component of the model is known as the *link function*, so that a logistic link function is being used in the linear logistic model.

4.1.2 Goodness of Fit test

The model selection procedure does not guarantee that the selected model closely represents all data. Another criterion is needed to check if the predicted probability from formula 4.1 is consistent with the data, if the vocalisation turns leading a PCD get higher probability than other vocalisation turns from this regression model. The goodness of Fit test is used to check the fitting quality.

The Hosmer-Lemeshow Statistic [Agresti, 2002] measures binary response model's Goodness of Fit, Chi-square distribution is assumed. Hosmer and Lemeshow recommend partitioning the observations into 10 equal sized groups according to their predicted probabilities. Then the statistics of difference should follow Chi-square distribution. If it significantly deviates from Chi-square distribution, the model's fitness is not accepted.

$$G^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi^2_8 \tag{4.4}$$

where

n_j : number of observations in the jth group

$O_j = \sum_i y_{ij}$: observed number of cases in the jth group

$E_j = \sum_i \widehat{p}_{ij}$: expected number of cases in the jth group

4.2 Supervised Learning Algorithms

In Section 3.1.2 I introduced the data formats and various features used in this study, and remarked that the instances for classification are highly imbalanced, as well as that the instances are not sampled independently but sequentially. These properties of the data set challenge the routines of classification, and this led us to evaluate the applicability of different classifiers.

Luz [2009] proposed conversational features based classification approaches for topic segmentation in a corpus of medical meetings, where Naïve Bayes classifier reached 27.6% accuracy (Pk), better than the approximately 40% accuracy[1]

[1] A smaller Pk score means a more accurate segmentation outcome.

(Pk) in lexical cohesion segmentation approach achieved from the AMI corpus [Hsueh and Moore, 2007a]. Naïve Bayes classification is one of our choices for AMI corpus topic segmentation, presented in Section 4.2.1. In addition, I assess classifiers which incorporate the relations between instances, instead of regarding each instance as an independent sample. Hidden Markov models (HMMs) and Conditional Random Fields (CRFs) are analysed in Section 4.2.3 to classify correlated VE instances. Finally, ensemble classifiers (e.g., Bagging and Boosting) are investigated as a way of solving the data imbalance problem, because they have the advantage of combining prediction from a group of classifiers.

4.2.1 Naïve Bayes

Bayes' theorem [Bishop, 1996] shows us the method to predict the class of an unknown instance given training instances. The posterior probability of a class labeling on the unknown instance is determined by the prior probability of classes, and the likelihood of attributes of each class. The likelihood $P(E|C_i)$ is calculated from training samples. Here C_i is the ith class, and E is the test instance.

$$P(C_i|E) = \frac{P(C_i)P(E|C_i)}{P(E)} \tag{4.5}$$

The standard Naïve Bayes classifier is based on the conditional independence assumption: within each class the values of the attributes from instances are independent. Consequently, $P(E|C_i) = P(v_1|C_i) \times \cdots \times P(v_a|C_i)$ where v_j is the value of the jth attribute in instance E. With NB, the posterior probability of class C_i on test instance E is shown in Equation 4.6. In topic segmentation, we only have two classes: *topic begin* instances and other instances. The predicted posterior probability is categorised with a threshold to assign class label for a test instance (e.g., $h(r) = 1$ if $P(C_i = 1|E = r) \geq \tau$). Binary NB classification is normally based on a Maximum *a Posteriori* (MAP) hypothesis where $\tau = 0.5$.

$$P(C_i|E) = \frac{P(C_i)}{P(E)} \prod_{j=1}^{a} P(v_j|C_j) \tag{4.6}$$

In practice, the independence is rarely satisfied. But results show that the

NB classifier is still optimal even when the independence assumption is grossly violated. In these experiments, Equation 4.6 may produce poor probability estimates, but the correct class will still have the highest estimate, leading to correct classification [Domingos and Pazzani, 1996]. NB tends to assign $P(C_i|E)$ extreme values, that is either, close to 0 or 1 [Elkan, 2001]. Zhang [2004] explained the good classification performance of NB from the sample dependency distribution and proposed that NB is optimal if the dependencies among attributes cancel each other out. NB is also capable of learning from both categorical and numerical features. For these reasons, as well as for comparison with previous work [Luz and Su, 2010], I choose NB as one of the techniques to be evaluated for topic segmentation.

4.2.2 Imbalanced data and thresholded Naïve Bayes classifier

In 30 AMI meeting files, the average quantity of vocalisation events (VE_c) in one meeting is 510, among which only 13.3 VEs lead a new topic (including both top-level and sub-level topics). Since the VE_c leading a new topic is defined as a positive instance for classification and the rest instances are defined as negative, the positive instances are only 2.6% of the data set. The highly imbalanced data set greatly challenges common classifiers. In many cases, Decision Tree, Zero-R and RBF Network classifiers for instance simply predict all instances as negative. The reason is that the instances of the positive class are not enough to train the classifier, being treated as outliers. Surprisingly, however, NB classifiers generate true positive predictions in many cases, due to its property introduced in Section 4.2.1. For example, Table 4.1 shows the confusion matrix of MAP NB[2] and unpruned Tree C4.5 (Min. 2 instances per leaf) with EP_{VOC} (n=3) feature set from meeting ES2002d.

In Table 4.1, there are 10 instances on topic boundary (positive class) and 777 as rest. 10-fold C4.5 classification only generates 4 positive predictions (0.5% of population), much lower than the actual portion of positive instances 1.3%, and

[2] Prediction $h(r) = 1$ if $P(C_i = 1|E = r) \geq 0.5$)

Table 4.1: Confusion Matrix of MAP NB(a) and C4.5(b) classification results with EP_{VOC} (n=3) feature set from meeting ES2002d. In this table, "a=1" and "b=0" refer to positive and negative instances in reference set respectively; "P" and "N" refer to predicted positive class and negative class respectively. MAP NB predicts 5 true positive instance, but C4.5 predicts no true positive instances.

(a)

P	N	← pred
5	5	a=1
79	698	b=0

(b)

P	N	← pred
0	10	a=1
4	773	b=0

all positive predictions are *false positives.* C4.5 suffers from data sparsity[3]. On the other hand, NB leads to 89.3% overall accuracy, although its precision is only 6% in positive predictions (5 *True Positive* predictions with 79 *False Positive* (FP) ones. We should notice that the predicted positive instances stand for 10.7% of the population, which is much higher than 1.3%, the true portion of positive instances. NB does not suffer much from insufficient training sets, but rather overpredicts. In this situation, reducing FP predictions will be an effective approach in improving segmentation performance.

MAP NB predicts an unknown instance as positive when its probability is higher than 50%. The existence of many FP cases indicates that it is too easy to make a positive prediction. Since the correct class will have the highest probability estimate [Domingos and Pazzani, 1996] (although the predicted probability is usually extreme), Luz [2009] proposed two ways in raising the threshold above 50% on positive prediction, in order to exclude some FP predictions. Specifically, one method is to filter the predictions with the ratio of boundary VE_c (positive class) within all VE_c, and only keep the positive predictions with the highest probabilities. This method is named as Proportional Threshold (*PT*) NB. The other method is to apply predefined probability threshold on predictions, and reserve the positive predictions with probabilities higher than a predefined threshold. This is named as Fixed Threshold (*FT*) NB. Both modifications of MAP NB will be tested in AMI topic segmentation experiments.

[3] A proper cost matrix could improve C4.5 classification accuracy, this approach will be conducted in future work.

4.2.3 Conditional random fields

In NB classification, feature values are assumed to be independent within one class, which simplifies the likelihood of a feature set as a product of the likelihood of each feature. Another assumption is that the order in which the features are presented is irrelevant to the calculation of the conditional probability. Although NB performs well, given the sequential nature of discourse, it seems necessary to assess a probabilistic model which accounts for sequential information among samples. As illustrated in section 3.1.2, each classification instance is a vocalisation event (VE), which is sampled sequentially from audio recordings. So ordinal relations naturally exist between instances, and make VE different from the randomly sampled instances used for most classifiers. Therefore, we hypothesise that a classification scheme incorporating time series relations should improve segmentation.

Hidden Markov models (HMMs) [Rabiner and Juang, 1993] represent a system as a Markov process with unobserved states, where a current state is not independent but is influenced by its preceding state. HMM has been successfully applied to speech recognition and other sequence labeling tasks. However, HMM is built on strong independence assumptions: each state depends only on its immediate preceding state, and each observation depends only on the current state. I seek a model with looser restrictions on independence. Conditional Random Fields (CRFs) [Lafferty et al., 2001] are undirected graphical models that encode a conditional probability distribution using a given set of features. CRF, as a discriminative model, specifies a conditional distribution $p(Y|X)$, which is conditioned on all observations X. This setting bypasses independence assumptions on features.

A hidden Markov model is encoded with joint probability distribution:

$$p(X, Y) = \prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t) \qquad (4.7)$$

where $X = Data$ and $Y = Labels$, the posterior probabilities of $Labels$ can be determined from joint distribution (generative model):

$$P(Labels|Data) = \frac{P(Data, Labels)}{P(Data)} \tag{4.8}$$

Equation 4.7 shows that HMM assumes dependence to exist only between two consecutive observations y_t and y_{t-1}. Compared to models that assume independent samples, HMM should be more appropriate for VE_c. However previous experiments of Luz [2009] indicate that the best P_k accuracy is observed at a level 3 vocalisation horizon, which means the current instance not only correlates to its closely preceding and following instances, but also correlates to discontiguous neighbour instances. Since CRF is a more flexible alternative to HMM, I do not investigate HMMs in topic segmentation. CRF, on the other hand, directly define the posterior distribution of labels through feature functions:

$$p(Y|X) = \frac{1}{Z(X)} \exp\{\Sigma_{t=1}^{T} \Sigma_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_{1:T}, t)\} \tag{4.9}$$

where $Z(X)$ is an instance specific normalization function. $Z(X)$ is introduced to guarantee that $p(Y|X)$ is a valid probability over all *Label* sequences.

$$Z(X) = \Sigma_y \exp\{\Sigma_{t=1}^{T} \Sigma_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, x_{1:T}, t)\} \tag{4.10}$$

In Equation 4.9 and 4.10, λ_k is a weight factor and $f_k(.)$ stands for feature function. In content-free topic segmentation, $t = 1, ..., T$ refers to a $Vocalisation\ Event$ (instance) in a sequence, and $k = 1, ..., K$ refers to each feature. So the conditional probability of an instance's label is proportional to an exponential sum over the weighted features of each instance.

The feature function $f_k(y_t, y_{t-1}, x_{1:T}, t)$ is defined upon a pair of adjacent labels y_t, y_{t-1}, the whole input sequence $x_{1:T}$ and the position of specific instance(s) (Figure 4.1). The output of feature function can be defined as binary. For example, a feature function $f_1(.)$ below is automatically generated from CRF model training step:

$$f_1(y_t, y_{t-1}, x_{1:T}, t) = \begin{cases} 1 & \text{if } y_t = \text{Boundary and} \\ & \quad x_{t+1} = \text{VOC3} \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

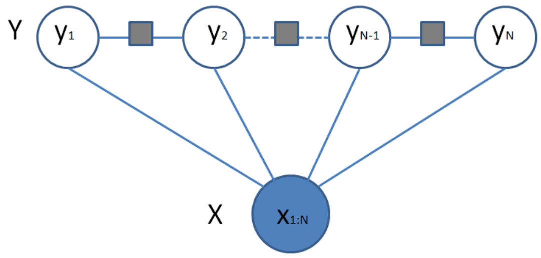


Figure 4.1: Linear-chain CRF

I assume that feature function $f_1(.)$ refers to feature *Vocalisation Duration*, then $f_1(.)$ is active when the *Label* of instance t is Boundary and the feature value of instance $t+1$ is VOC3. The effect of feature *Vocalisation Duration* depends on its corresponding weight λ_1. If $\lambda_1 > 0$, whenever f_1 is active, it increases the probability of the *Label* sequence $y_{1:T}$. If $\lambda_1 = 0$, f_1 has no effect on the conditional probability. CRF training means to find the λ parameters in a CRF [Zhu, 2010].

4.2.4 Imbalanced data and ensemble classifiers

Ensemble classification [Hastie et al., 2009] is a supervised learning scheme which combines the predictions of multiple classifiers. As a consequence, ensemble classifiers produce collective decisions, which are more robust than a single classifier in many situations. Bagging and Boosting are typical ensemble classifier techniques. Bagging predictors are methods for generating multiple versions of a predictor from one training set by performing a plurality vote when predicting a class, where the multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets [Breiman, 1996].

Boosting was introduced by Freund and Schapire [1997] to produce one accurate prediction by combining moderately inaccurate predictions from a group of weak learners. The most popular Boosting algorithm is AdaBoost. While Bagging relies on *random* and *independent* changes in the training data implemented by bootstrap sampling, Boosting advocates *guided* changes of the training data to direct further classifiers toward more "difficult cases" [Kuncheva et al., 2002]. AdaBoost highlights two main steps distinguishable from other classification algorithms and constrains the training procedure under "guidance". The first step is to run a base learner repeatedly for T times, and to maintain a distribution (a set of weights) over the training set (e.g., the weight on training example i on round t is $D_t(i)$). The second step is to update the weights in each round. Initially, all weights are set equally, but on each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set [Freund and Schapire, 1999]. The final prediction is the categorised weighted sum of the predictions from base learners.

Revisiting Table 4.1, I expect AdaBoost to highlight the FP and FN instances (hard cases) and lay minor weight on the majority of data set: True Negative (TN) instances (easy cases).

Chapter 5

Evaluation Methodologies

For standard categorisation tasks, *Precision*[1] and *Recall*[2] [Bishop, 1996] are metrics commonly used to judge how well the classification result matches the reference, regardless of order. In contrast, segmentation methods are designed to locate segment boundaries along a sequence of instances. A "good" segmentation algorithm should have similar quantities of segments compared to reference segmentation, and begin/end boundary positions should match with reference.

If we solve a segmentation problem with categorisation methods, it is mandatory to specify the number of categories. Since it is hard to categorise all kinds of topics with text-independent features, I prefer to convert the segments (each includes some instances, e.g. sentences or vocalisation events) into a binary labeled data set: segment boundary instances and non-boundary instances. Under this setting, if the predicted boundaries are missing, or more than necessary, it is natural to represent the loss by *false negatives* (FN) or *false positives* (FP) predictions respectively, and further by *Precision* and *Recall.* However, with fixed precision and recall scores (or number of wrong predictions in each class), the distribution of boundary instances may vary, and produce major differences in terms of goodness of segmentation.

There is a need for alternative metrics to evaluate segmentation fitness. Such metrics should favor a prediction with same number of instances as reference,

[1] Precision is the fraction of correctly identified true instances out of all identified true instances.

[2] Recall is the fraction of correctly identified true instances out of all true instances.

in either class. Furthermore, upon one mistaken prediction such metrics could measure its distance to the right sample, and punish less on a *near-miss* error. As a conclusion, categorisation assumes independence of samples and counts only the number of correct predictions. Segmentation through categorisation should consider the sequential relation between samples. Therefore metrics developed in text segmentation are adopted in content-free topic segmentation, such as P_k and WD (Section 5.1).

5.1 Metrics

Research in topic-based text segmentation brings valuable examples of evaluation methods. Beeferman et al. [1999] proposed an error metric P_k for text segmentation, where k is equal to half of the average sentence quantities in a reference topic segment. Then the number of mismatch is counted by a moving window with $2k$ sentences. At each sentence break, P_k algorithm checks whether the two ends of the window are in the same segment in the reference segmentation, and increase a counter if the result from automatic segmentation disagrees. Finally the counter sum is divided by the number of total iterations, and results in a scalar between 0 and 1. This is P_k. A higher P_k value corresponds to a worse match. In this study, P_k is used to assess content-free topic segmentation, and the degree of mismatch is counted on vocalisation events instead of sentences. Moreover, the commonly used classification concepts $False\ positive$ (FP) and $False\ negative$ (FN) are updated for content-free topic segmentation. FP is defined as an extra segment boundary identified by experimental algorithm, but not in reference segmentation. FN is a missed segmentation boundary by experimental algorithm.

The P_k metric is welcomed by researchers, and is utilised in the study of content-free topic segmentation. However, Pevzner and Hearst [2002] found some drawbacks in P_k metric:

1. False negatives get penalized more than false positives. In the ideal case, all segments are of same length $2k$. Then every false negative error gets penalty k. But for false positive errors, when the extra boundary is near (with distance $\leq k$) to previous or next reference boundary, the penalty is

$\leq k$.

2. Number of boundaries is ignored

3. *Near-miss* errors get penalized too much (*near* means within a distance of $\frac{k}{2}$)

In order to avoid these reservations, Pevzner and Hearst [2002] proposed *WindowDiff* (WD) (Equation 5.1) as an alternative metric for segmentation. In this equation, $b(i, j)$ represents the number of boundaries between position i and j in the text and N is the number of sentences in the text.

$$WindowDiff(ref, hyp) = \frac{1}{N-k} \sum_{i=1}^{N-k} (|b(ref_i, ref_{i+k}) - b(hyp_i, hyp_{i+k})| > 0) \tag{5.1}$$

For each position of a window with k units[3], WD metric compares how many reference segmentation boundaries fall in the interval (r_i) and how many boundaries are assigned by segmentation algorithm (a_i). The segmentation algorithm is penalized if $r_i \neq a_i$. WD has the advantage of avoiding distortions caused by the imbalance between FP and FN which characterises topic boundary spotting.

In Section 5.3, I discuss a weakness of P_k and WD with respect to under-predicted cases, and introduce a supplement metric *balance factor* ω to monitor the quantity of predicted boundaries.

5.2 Near-miss errors

In speech segmentation, when I have a *near-miss* error (the segmentation boundary from automatic algorithm is near to reference boundary), meeting audience will take some effort to relocate the beginning of a segment. A *near-miss* does not harm segmentation results much, but a $missing$ boundary or $superfluous$ boundary has a much worse influence, since the index of segmentation is tangled, and audience has much trouble to locate the true position of the asked segment.

[3] The window is defined in the same way as P_k

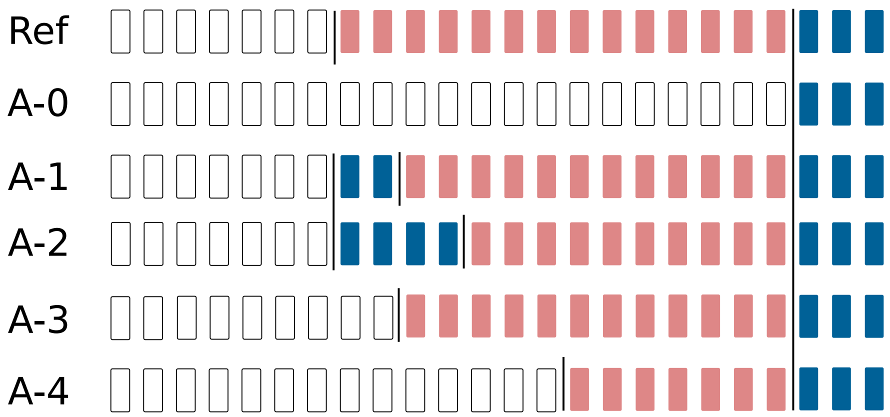


Figure 5.1: A reference segmentation and five different hypothesised segmentations with different properties. (Adapted from Pevzner and Hearst [2002])

It is fair that the selected metric punishes more on the *missing* and *superfluous* boundaries, but less on *near-miss* cases.

An example from Pevzner and Hearst [2002] clearly compares the difference between P_k and $WindowDiff$, on the punishment of *near-miss* error.

In figure 5.1, A-4 is the worst case, it contains both false positive and false negative errors. A-0 has a false negative error, and A-2 has a false positive error. The error in A-3 can be regarded as *near-miss* error (note the error quantity as e), because the mis-labeled units are within $\frac{k}{2}$ distance. The error quantity in A-1 is only e, and A-1 correctly locates two boundaries, but it contains an false positive error. In PCD segmentation, A-1 is worse than A-3. A superfluous segment harms the segment index, while a *near-miss* error does not.

When P_k metric is used to evaluate the errors in these cases, A-4 gets penalty $2k$ if the false positive boundary falls in the middle of reference boundary. A-0 gets penalty k, and A-2 gets penalty $\leq k$. The penalty for A-1 is e, while the penalty for A-3 is $2e$. This is not true in speech segmentation, A-1 should be punished more than A-3.

$WindowDiff$ metric performs differently from P_k. A-4 gets penalty about $2k$, and A-0, A-1, A-2 get same penalty, about k. But A-3 gets penalty of only $2e$, where e is assumed to be much smaller than k, so a *near-miss* is penalized less than false positive and false negative. This character is coherent with the needs in PCD segmentation, so $WindowDiff$ is favored. In the end, a weak point in

$WindowDiff$ should be mentioned: A-1 and A-2 get same penalty, while the actual quantity of error is not the same. In PCD segmentation, the consequence is that the more extra 'pieces' in segmentation result, the heavier it is penalized. But the positions of superfluous boundaries are not considered. This phenomenon is not influential in primary PCD segmentation algorithm design, but it will be a bottleneck on the 'fine tuning' of algorithm.

5.3 Weakness of popular segmentation metrics

As discussed in the beginning of Chapter 5, the standard classification metrics $Precision$ and $Recall$ evaluate classification instances as independent samples, so they cannot judge the sequence information in segmentation. In other words, these two metrics precisely measure the effects of missing and surplus boundary predictions, but they do not evaluate the adjacency of predictions (counted as the numbers of VE mismatch) to reference positions (the manually labelled topic boundary positions). So, $Precision$ and $Recall$ are not proper choices to evaluate the goodness of segmentation.

Within the AMI corpus (introduced in Section 3.1), I apply the well established segmentation metrics P_k and WD in content-free topic segmentation experiments. Although P_k and WD calculate the degree of mismatch in different methods (Section 5.1 and 5.2), they approach 0 in case of perfect match between prediction and reference, and approach 1 for erroneous predictions. In Section 6.2, the segmentation accuracy of several classifiers are evaluated with P_k and WD, which represent the relative goodness of each classifier.

So far P_k and WD are assumed to be perfect for segmentation evaluation, but when I revisit in Chapter 6 the confusion matrix of Conditional Random Fields (CRF) and naïve Bayes classifiers (Table 6.7), the deficiency of P_k and WD are uncovered. In Table 6.6 (page 81), CRF classifier presents a better WD score than proportional threshold naïve Bayes classifier (PT NB, introduced in Section 6.2.4) with each feature set, but the classification confusion matrix (Table 6.7, page 82) shows that CRF generated many fewer boundaries than PT NB. A topic boundary prediction scheme with very few boundaries offers little help to meeting audience to locate reference points. So, a better P_k or WD score does not guarantee a better segmentation result. A new metric is needed to evaluate segmentation outcomes impartially.

In order to ascertain the existence of discrepancy between P_k, WD and the number of boundary predictions, I test the extreme situations of no boundaries, all boundaries, and boundaries distributed uniformly or randomly along vocalisation events where the number of boundaries are identical to reference. The segmentation accuracy is listed in P_k and WD (Table 5.1).

No boundaries and *All boundaries* are the two worst situations for segmentation, but their P_k and WD differ greatly (row 1 and row 2 in Table 5.1). In perspective of segmentation usefulness, *Uniform boundaries* are better than either *No boundaries* or *All boundaries*, but their $P_k =$ and WD scores (row 3 and row 4 in Table 5.1) fall in between *No boundaries* and *All boundaries*. Comparing baseline scores and classifier scores, *no boundaries* ($WD = 0.405$) gives a better WD score than PT NB ($WD = 0.429$ with 59 true positive predictions) on EP_{GAP} (Table 6.3 and 6.7(d)). These facts disclose main drawbacks of P_k, WD:

1. *No boundaries* and *All boundaries* are equally ineffective for the retrieval of segments, but they are not punished on the same scale.

2. A better P_k or WD score does not guarantee a better segmentation result.

3. P_k and WD do not establish a linear trend between the perfect match and the worst cases.

Since segment boundary markers are sought to offer reference points for browsing, it would be misleading to choose a classifier with very few boundary predictions. Both P_k and WD are not effective to gauge the effect of boundary quantity, so it is necessary to consider additional metrics. The confusion matrix indicates the number of true boundaries ($TP+FN$) and predicted boundaries ($TP+FP$), and distinguishes a useful segmentation from a trivial one. In order to present the figure in a straightforward way, I propose a concise alternative to confusion matrix which can be reported alongside P_k and WD results. It consists of calculating what I call *balance factor* ω in the manner shown in equation (5.2).

$$\omega = \frac{No.\ of\ Predicted\ Boundaries}{No.\ of\ Real\ Boundaries} = \frac{TP + FP}{TP + FN} \tag{5.2}$$

A perfect segmentation scheme should present the same quantities of segments or boundaries as the reference, where $\omega = 1$. Deficient boundaries correspond to $\omega < 1$ and surplus boundaries correspond to $\omega > 1$. For extreme cases, a *no boundaries* segmentation corresponds to $\omega = 0$, and for the *all boundaries* case $\omega \gg 1$. In case a classifier generates 0 TP, m FP and m FN, then $Precision = 0$ and $Recall = 0$, but $\omega = 1$. These do not contradict a good segmentation with $P_k \approx 0$ and $WD \approx 0$ where same quantity of boundaries are predicted and each is located next to a true boundary. ω distinctively examines the effect of the quantity of predicted boundaries, while P_k and WD evaluate the quality of segmentation, thus complementing each other. Table 5.1 shows that *uniform boundaries* and *random boundaries* have $\omega = 1$, much better than *no boundaries* and *all boundaries*. Thus ω establishes a linear scale between good predictions and bad ones, which compensates the three drawbacks of P_k and WD. A prediction result with $\omega \approx 1$ and $P_k \approx 0$, $WD \approx 0$ is a useful set of indicators for topic segmentation.

Table 5.1: Segmentation accuracy scores P_k, WD and ω in 4 baseline conditions. Test set is a concatenated data set of 30 meetings. The first row in table shows the accuracy scores when no VE is topic boundary in all dataset. The second row shows the accuracy scores when each VE is a topic boundary. The third row stands for the case that the number of predicted topic boundaries is same as the number of true boundaries, and the predicted boundaries are uniformly distributed. The fourth row stands for the same situation as the third row except that the predicted boundaries are randomly distributed. The fifth row stands for the situation that the predicted boundaries have exactly the same quantity and positions as true boundaries.

	P_k	WD	ω
No boundaries	0.389	0.405	0
All boundaries	0.611	1.0	34.3
$*$Proportional uniform boundaries	0.506	0.565	1.0
$*$Proportional random boundaries	0.477	0.548	1.0
Perfect match	0	0	1.0

$*$ where the number of boundaries is same as reference.

5.4 Receiver Operating Characteristics

In Section 5.3, *balance factor* ω is highlighted in complementing segmentation metrics P_k and WD, but since the deficiency of segmentation metrics comes from the lack of positive outcomes in topic boundary classification, successful classification metrics have the potential in detecting unbalanced classification outcomes. In this section, a popular classification performance measure is studied and is compared with *balance factor* ω, in order to access the novelty and necessity of ω.

Receiver Operating Characteristics (ROC) graph is a technique for visualising, organising and selecting classifiers based on their performance [Fawcett, 2006]. There are many metrics indicating a classifier's performance, such as true positive rate, false positive rate, precision, recall and accuracy. Among these metrics, ROC graph highlights the relation between false positive rate and true positive rate, which are illustrated in Figure 5.2, Equation 5.3 and Equation 5.4.

		Reference classes	
		+	**-**
Predicted classes	**Y**	**T**rue **P**ositive	**F**alse **P**ositive
	N	**F**alse **N**egative	**T**rue **N**egative
Column total:		**Pos**	**Neg**

Figure 5.2: Confusion matrix

Figure 5.2 presents the cross table of all four possible outcomes of classifying an instance. If the instance is positive(+) and is classified as positive(**Y**), it is counted as a *True Positive*; if it is classified as negative(**N**), it is counted as a *False Negative*. If the instance is negative(**-**) and it is classified as negative(**N**), it is counted as a *True Negative*; if it is classified as positive(**Y**), it is counted as a *False Positive*. Then FP rate is counted as the ratio of FP instances out of all negative instances (Equation 5.4), and TP rate is counted as the ratio of TP instances out of all positive instances (Equation 5.3), which is identical to *Recall* (Equation 5.6).

$$TP\ rate = \frac{TP}{Pos} = \frac{TP}{TP + FN} \tag{5.3}$$

$$FP\ rate = \frac{FP}{Neg} = \frac{FP}{FP + TN} \tag{5.4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.5}$$

$$Recall = \frac{TP}{Pos} = \frac{TP}{TP + FN} \tag{5.6}$$

$$Accuracy = \frac{TP + TN}{Pos + Neg} \tag{5.7}$$

Recall my definition of a balance factor ω from Equation 5.2, repeated here in Equation 5.8, none of the metrics in Equation 5.3 to 5.7 stands for the identical measurement as it does. However, ROC graph integrates TP rate and FP rate in evaluating classification errors. It is important to measure the relation between ROC graph and ω, and reach a conclusion whether ω could be replaced by ROC graph in segmentation evaluation.

$$\omega = \frac{No.\ of\ Predicted\ Boundaries}{No.\ of\ Real\ Boundaries} = \frac{TP + FP}{TP + FN} \tag{5.8}$$

An ROC graph is a two dimensional space where X axis is FP rate and Y axis is TP rate. In this space, the performance of a classifier is either represented by a single point, or a polyline. *Discrete* classifiers (such as Decision Tree), only output a class label for each instance. As a consequence, a discrete classifier is represented as a single point on ROC graph, with TP rate and FP rate derived from confusion matrix. When comparing two discrete classifiers, the one located upper left to the other one in ROC graph is better.

The other type of classifiers, *probabilistic* classifiers (such as naïve Bayes classifier), output the probability value or membership score of a class on each instance. A binary classification result is generated by applying a threshold over probability values. For each instance, if its classification output is above the threshold, it is classified as positive(**Y**), otherwise negative(**N**). For each threshold value, the classifier generates one confusion matrix as Figure 5.2 and locates one point on ROC graph. If all choices of threshold (from $-\infty$ to $+\infty$) are examined, an ROC curve could be plotted through the trace of all points. In practice, it is not efficient to enumerate all threshold values under a fixed granularity. Fawcett [2006] proposed an algorithm of calculating (FP, TP) pairs based on a sorted list of output probabilities. A sample ROC curve is shown in Figure 5.3. When an ROC curve is drawn, the area under curve (AUC) is used to indicate the average performance of classifiers. A higher AUC score stands for better performance.

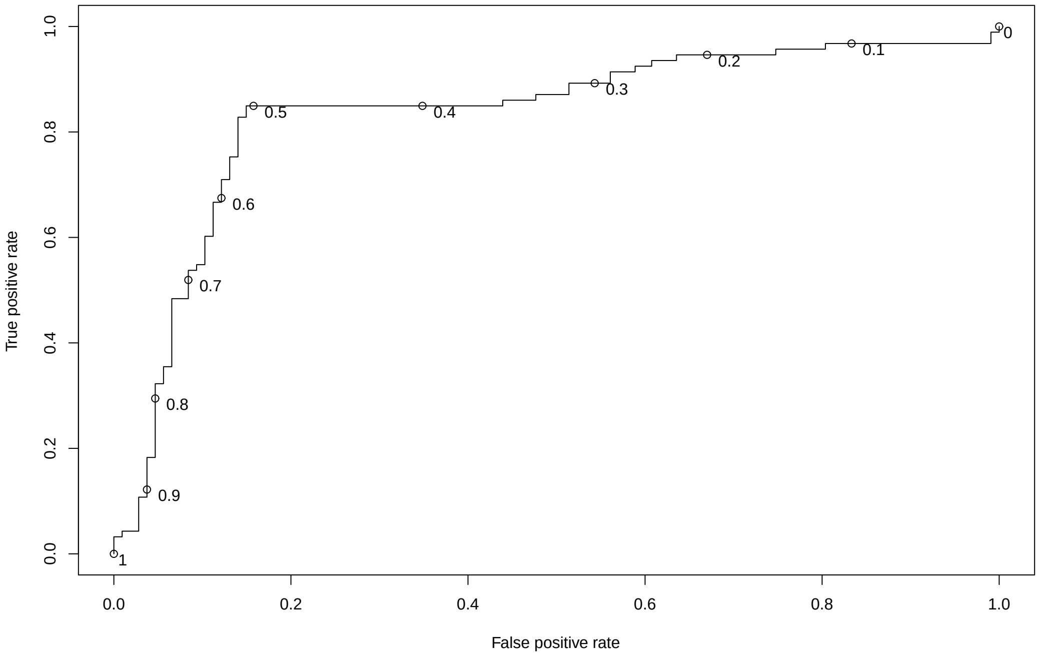


Figure 5.3: A sample ROC graph with probability thresholds

Another important characteristics of ROC space is *iso-performance line* [Provost and Fawcett, 1998]. If two points A (FP_1, TP_1) and B (FP_2, TP_2) are on the same *iso-performance line*, they have the same performance (or expected classification cost). The expected overall classification cost of one classifier is composed of its false positive and false negative costs:

$$\begin{aligned} cost &= p(Pos) \cdot FN \cdot c(N, Pos) + p(Neg) \cdot FP \cdot c(Y, Neg) \\ &= p(Pos) \cdot (1 - TP) \cdot c(N, Pos) + p(Neg) \cdot FP \cdot c(Y, Neg) \end{aligned} \tag{5.9}$$

In Equation 5.9, $p(Pos)$ and $p(Neg)$ stand for the portion of positive and negative instances in reference set; $c(N, Pos)$ and $c(Y, Neg)$ stand for the cost of false negative and false positive predictions, respectively. If two classifiers A and B have the same overall cost, $cost_A = cost_B$, then Equation 5.9 is transformed into Equation 5.10, where A and B are on the same *iso-performance line* with slope m. When m is defined, I can move *iso-performance line* from upper left to lower right of the ROC space, the first point (classifier) on line has better performance than the rest. As a consequence, the *iso-performance line* is convenient tool in evaluating classifiers in ROC space.

$$\frac{TP_2 - TP_1}{FP_2 - FP_1} = \frac{p(Neg)c(Y, Neg)}{p(Pos)c(N, Pos)} = m \tag{5.10}$$

5.4.1 ROC and balance factor

Since ω is only related to TP, FP, FN, it is interesting to know whether ROC graph has similarity with ω in evaluating segmentation outcomes, and could replace ω. If it is, popular classification metrics could also be used to gauge segmentation outcomes.

In order to test this hypothesis, segmentation experiments in the AMI corpus (Section 6.2)are reused here. MAP Naïve Bayes classifier (MAP-NB) and Fixed Threshold naïve Bayes classifier (FT-NB) are tested upon four feature sets in the AMI corpus. Table 5.2 contains results from Table 6.2 and Table 6.3.

$$\omega = \frac{TP + FP}{TP + FN} = \frac{TP + FP}{Pos} = (TP\ rate) + \frac{FP}{Pos} \tag{5.11}$$

Figure 5.4 shows ROC graphs of naïve Bayes classifier on four AMI feature sets. A ROC curve is generated from NB classifier predicted probabilities on each instance and flexible positive class thresholds among [0,1]. When a probability threshold is determined and is applied to each instance, a classifier's outcome is

Table 5.2: Segmentation accuracy of MAP and Fixed Threshold Naïve Bayes classifier (Adjacent boundaries are removed)

		P_k	WD	ω
MAP-NB	EP	0.408	0.54	1.78
	EP_{VOC}	**0.365**	**0.496**	1.51
	EP_{VOCP}	0.405	0.558	1.92
	EP_{GAP}	0.408	0.599	2.48
FT-NB$*$	EP	0.341	0.41	0.553
	EP_{VOC}	0.328	**0.406**	0.654
	EP_{VOCP}	0.34	0.423	0.716
	EP_{GAP}	**0.326**	0.44	1.144

$*$ threshold is 99% for positive class

binary and the classifier's performance is represented as one point in ROC space. The point 0.5 on each curve stands for MAP NB classifier (positive class threshold is 50%), and the point 0.99 stands for FT-NB classifier (positive class threshold is 99%). Since it is hard to compare classifier performance in 4 sub-figures, I re-organise binary classifiers in Figure 5.5.

Because I do not have the relative cost between a missing topic boundary (false negative) and a redundant boundary (false positive), FP and FN are assumed to cost equally in Equation 5.10. On the other hand, positive instances (topic boundaries) only stand for about 1% instances in corpus. So, the slope of *iso-performance line* is $m = 99$. In Figure 5.5, both classifiers perform best with EP, which contradicts ω in Table 5.2. For FT-NB, EP feature set has $\omega = 0.553$, which means only half of the boundaries are predicted. EP is inferior to EP_{GAP}, which owns $\omega = 1.144$. As a consequence, ROC measures indicate classification performance, but do not satisfy segmentation need, and do not replace balance factor ω. Equation 5.11 shows that, ω cannot be represented by TP rate and FP rate. So, ω will be used as a supplementary segmentation metric, instead of ROC measures.

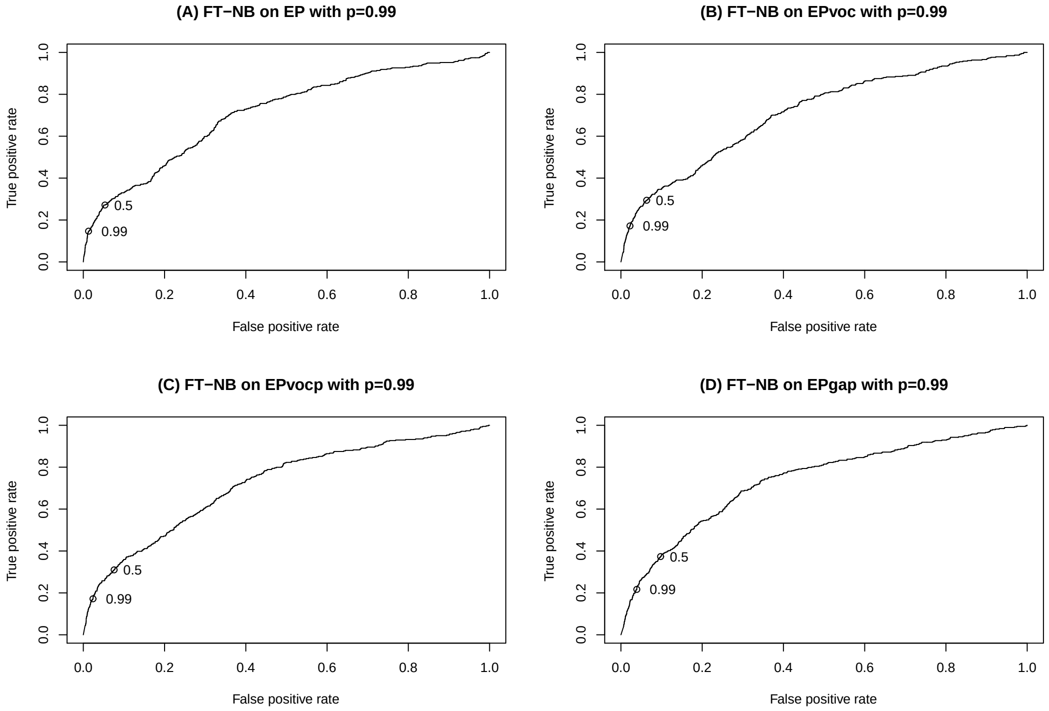


Figure 5.4: ROC graph of naive Bayes classifiers on 4 feature sets, MAP-NB and FT-NB correspond to p=0.5 and p=0.99 respectively

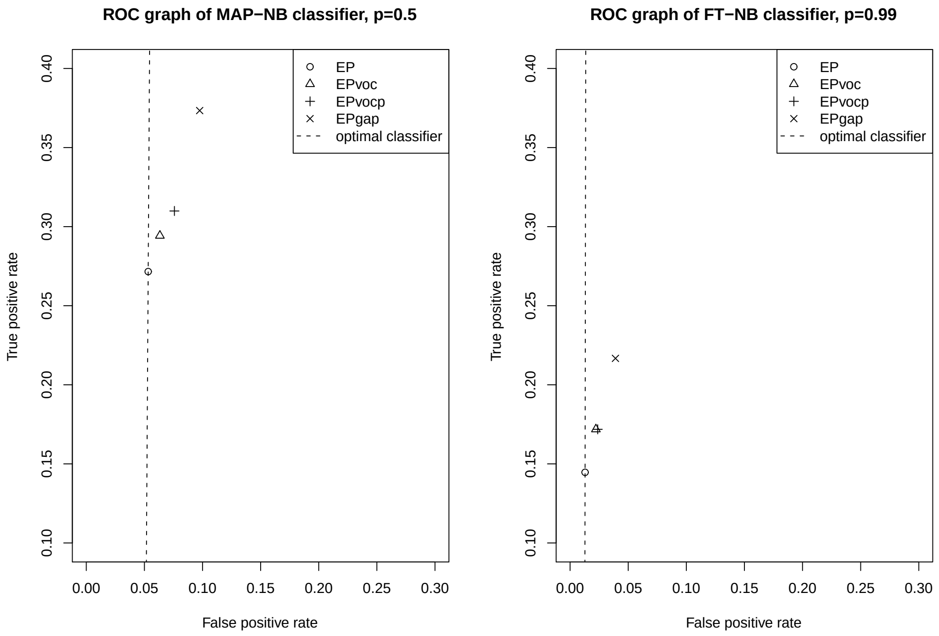


Figure 5.5: *iso-performance line* (slope $m = 99$) and discretised naive Bayes classifiers

Chapter 6

AMI Experiments

Chapter 4 introduced various classification algorithms and proposed vocalisation event as the unit representation for audio analysis, instead of fixed length sequential audio samples. In this chapter I present a series of experiments with the AMI corpus, including content-free topic segmentation and meeting phase effect. The primary concern is to assess the effectiveness of topic segmentation with the proposed classification algorithms and data structure. Beyond which, an essential procedure is to select proper metrics for topic segmentation. I will discuss the drawbacks of popular metrics and propose a method to amend them (Section 5.3).

In classification experiments (Section 6.2), I compare the performance of various classifiers including decision tree (Section 6.2.2), naïve Bayes (Section 6.2.3 and 6.2.4), CRF (Section 6.2.5) and ensemble classifiers (Section 6.2.6). In order to verify vocalisation horizon effect and find more influential features, I compare the effectiveness of the above methods using empty pause and full pause (Section 6.3.1), as well as speaker role horizon (Section 6.3.2) as topic segmentation features. Finally, meeting phase effect is evaluated in Section 6.4.1.

6.1 Data set

Since classification approaches are proposed for content-free topic segmentation, a clear definition of classification sample is mandatory. Vocalisation Event (*VE*) is introduced in chapter 3.1.2 as a piece of continuous speech from one speaker without an empty pause longer than 0.5 second. *VE* is identified sequentially from audio stream by its start and end time, hence *VE* indicates a piece of talk from one speaker, and it embodies many vocalisational characters[1]. Such characters are integrated as classification features. As a comparison with *VE*, audio samples from a fixed length window cannot represent complete vocalisation characters from a speaker, and also cannot represent vocal interactions among speakers precisely, because vocalisation events are intercepted by sample windows.

6.1.1 Vocalisation Horizon

Vocalisation Horizon (*VH*) is a novel feature coined for topic boundary *VE* classification. Most classifiers assume independent samples and abandon context information from preceding and following *VE*. *VH* then is employed to express context information and is integrated as a feature of the current *VE*.

Theoretically, any feature from adjacent *VE* can be used as *VH* of the current *VE*. Vocalisation duration, pause or overlap duration, speaker ID or speaker role, average pitch, volume or talking speed, etc, all of these are potential choices. In this work, the basic feature *vocalisation duration* is sought for *VH* as the first choice. The reason is that, at the boundaries of sentences or topics, speakers generally lengthen the last syllables or phones, since they slow down the speaking rate [Shriberg et al., 2000]. This phenomenon is well established with phones, but in our content-free design the highest resolution is *VE* instead of phone, so it is difficult to gauge the variation of specific phone duration. Alternatively, I emphasise the variation of *VE* duration as a classification feature. Vocalisation Event duration is not only used as a classification feature, but is an element of vocalisation horizon.

In addition to VE duration, there are more vocalisation features suitable to be elements of vocalisation horizon. As indicated by Esposito et al. [2007], empty pauses and filled pauses are used as a linguistic means for discourse segmentation. A long pause and a short pause demonstrate different dialogue structure as well as speakers' intention. Figure 3.1 shows that, pause and overlap[2] among adjacent

[1] Vocalisational characters used in this study are defined in Equations 3.1 to Equations 3.8

[2] Overlap is the phenomenon that the second speaker starts to talk before the first speaker

VEs are defined as pause horizon and overlap horizon, which reflect the pace of group interaction. I expect pause/overlap horizon could be used in combination instead as single features to better detect topic boundaries. The same is true for *VE* horizon, the duration of pause/overlap is the only feature used from them, and the level of horizon regulates how many adjacent pause/overlap are sampled. Section 6.3.1 illustrates the methods to extract empty pauses and filled pauses separately.

Acoustic features are well studied in discourse structure analysis, and I also propose some of them as options of *VH*. Shriberg et al. [2000] presented solid steps in pitch modelling and used pitch slope and range as features. Due to practical constraints, I leave out complex procedures of pitch manipulation and only integrate average pitch intensity in a *VE* as an acoustic feature. Pitch intensity is also used as an element of horizon. Prosody brings even greater difficulties than pitch. I instead use *word rate* in a *VE* to stand for prosody.

6.1.2 Features and settings

As introduced in section 3.1.2, various features can be employed in content-free topic segmentation, such as vocalisation duration, pause and overlap duration, speaker role, pitch, talking speed, and their corresponding horizon format. For convenience of classification method comparisons in this chapter, I only use a simplified feature set generated without considering filled pauses. Specifically, the features are represented as Equations (3.5) to Equations (3.8).

30 meetings from the ES section [Carletta, 2007] of the AMI corpus are included in this study. For each experiment, instances from 30 meetings are concatenated[3], filename being added as an extra nominal feature[4]. I run a batch of 30-fold cross validation for each classifier, and in each fold, instances are collected sequentially to preserve order.

stops talking. The duration of overlap is recorded from the time that the second speaker starts talking to the time that the first speaker stops talking. The duarion of overlap is the only feature used from overlap in this study.

[3] For a single meeting, there are less than 20 topics in general. As a consequence, it is very difficult to train classification models with N-fold cross-validation. However, in a concatenated dataset, the quantity of boundary instances is much higher, and facilitates classifier training.

[4] same data setting is used for all experiments in this study unless otherwise indicated.

In order to increase segmentation accuracy, classification outputs are post processed through a filtering step. A binary filter removes any *consecutive* topic boundaries predicted, except the first one. This filtering method is named as **NoAdj**. The reason of filtering is that, in classification every instance is treated as independent sample and there is no sequential relations among instances. However, instances in segmentation are sequentially placed and two consecutive boundaries are erroneous. A simple solution is to reserve only the first boundary predicted, although there are other filtering strategies[5].

6.2 Classification schemes for topic segmentation

Two aspects are essential for content-free topic segmentation: classification method and feature set selection. I study these two aspects separately and introduce classification schemes in this section. Content-free topic segmentation has definite fitness criteria (Chapter 5), different from classification accuracy. In this section, classifiers are evaluated with segmentation metrics. Moreover, classifier performance with correlated and unbalanced samples is of special concern. Proper classifiers need to accommodate correlated instances and highly unbalanced dataset in two classes. A naïve Bayes classifier generates robust predictions with inaccurate probability estimation [Domingos and Pazzani, 1996], and it is optimal if the dependencies among attributes cancel each other out [Zhang, 2004]. Conditional Random Fields (CRF), as a discriminative model, specifies conditional probability on all observations and precludes the independence assumption [Lafferty et al., 2001]. Ensemble classifiers are designed to learn a more expressive concept than a single classifier. These classifiers are assessed upon AMI data for successful topic segmentation solutions. A decision tree is a typical classification model constructed on information gain, I use decision tree segmentation results as reference to evaluate other classifiers.

[5] For example, I could reserve the instance with the highest posterior probability, in a probability based classifier

6.2.1 Objectives

The initial step of applying classification approaches to content-free topic segmentation, is to validate the appropriateness of reducing segmentation as classification. Vocalisation events are categorised into two classes: topic boundary VE and other VE, which enables binary classification schemes for segmentation. Segmentation accuracy is assessed in terms of P_k, WD and ω. Based on all of these preparations, the most essential step is to find proper classifiers and feature sets to fulfill the design. I hypothesise that, some classifiers are insensitive to correlated and unbalanced samples, and generate satisfying accuracy on topic boundaries. The tasks of the experiments are summarised as following objectives:

Objective 1: Discover which classification schemes are applicable for content-free topic segmentation, and deliver acceptable segmentation accuracy.

Objective 2: Evaluate various vocalisational features and determine which features with which form are suitable for segmentation.

In order to validate this hypothesis, I employ a common data set with limited features in section 6.1.2.

6.2.2 Decision trees

First I use the well established C4.5 decision tree as a classifier to distinguish topic boundary *VE*s from others. Unpruned trees are employed with minimum number of instances per leaf set to $M = 5$. 30-fold cross-validation is performed on the concatenated data set of 30 AMI meetings.

Table 6.1: Segmentation accuracy of C4.5 decision tree on four feature sets with empty pause settings

Feature Set	P_k	WD	ω
EP	0.49	0.643	2.49
EP_{VOC}	0.477	0.623	2.38
EP_{VOCP}	0.467	0.621	2.4
EP_{GAP}	0.475	0.634	2.53
$*$ Baseline	0.506	0.565	1

$*$ accuracy of proportional uniform boundaries

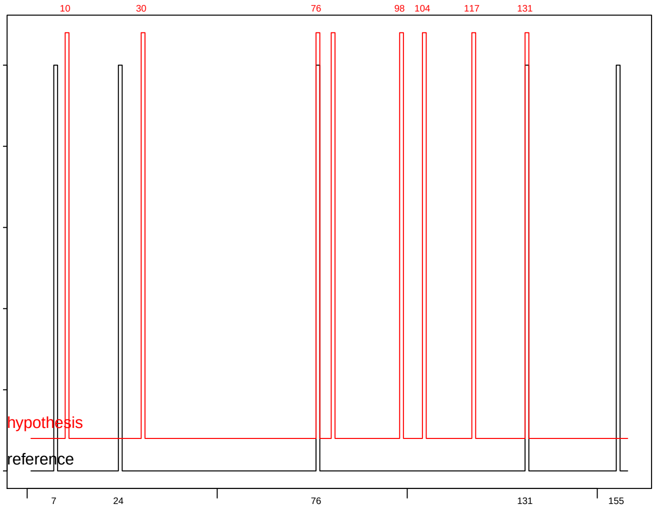


Figure 6.1: C4.5 prediction with EP_{VOCP} features. "reference" indicates the manually assigned topic boundaries and "hypothesis" stands for the predicted boundaries from C4.5 classifier.

EP_{VOCP} generates the best segmentation accuracy for C4.5 unpruned decision tree, part of its boundary predictions are shown in Figure 6.1. We can see that C4.5 has good prediction accuracy for reference positions. VE No. 76[6] and 131 match reference, and VE No. 10, 30 are close to reference No. 7, 24. However, C4.5 produces too many false positive cases, which is confirmed by $\omega = 2.4$. A high ω score is common with C4.5 on all feature sets. Comparing C4.5 segmentation accuracy with *random boundaries* accuracy (Table 5.1), I see that *random boundaries* has better score on each metric (if the number of boundaries is known). So C4.5 decision tree is not a promising classifier for content-free topic segmentation. Its surplus predictions obscure audience from locating the right boundaries.

6.2.3 Naïve Bayes classifier

The naïve Bayes classifier is well known for its capability to simplify posterior probability density estimation by a product of marginal distribution of features given their class label. Although this assumption is challenged by many non-

[6] Vocalisation Event is the only unit in segmentation, instead of time and frame. No.76 refers to the 76th VE in the corresponding meeting record.

independent feature sets, Naïve Bayes often generates satisfying predictions. *VE* features are correlated, especially horizon features. Since decision tree suffers from these features, I am interested in Naïve Bayes classifier, a probability based model.

Maximum a posteriori (MAP) is a classical rule of the naïve Bayes classifier. The class with the highest posterior probability estimation will be assigned to the instance as prediction. For topic boundary classification, MAP means the class with higher than 50% probability will win. Same as in Decision Tree experiments (Section 6.2.2), four *EP*s features are involved in NB experiment. 30 fold cross validation produces better segmentation performance than C4.5. The output is in Table 6.2.

Table 6.2: Segmentation accuracy of MAP Naïve Bayes classifier on four feature sets with empty pause settings (with $NoAdj$)

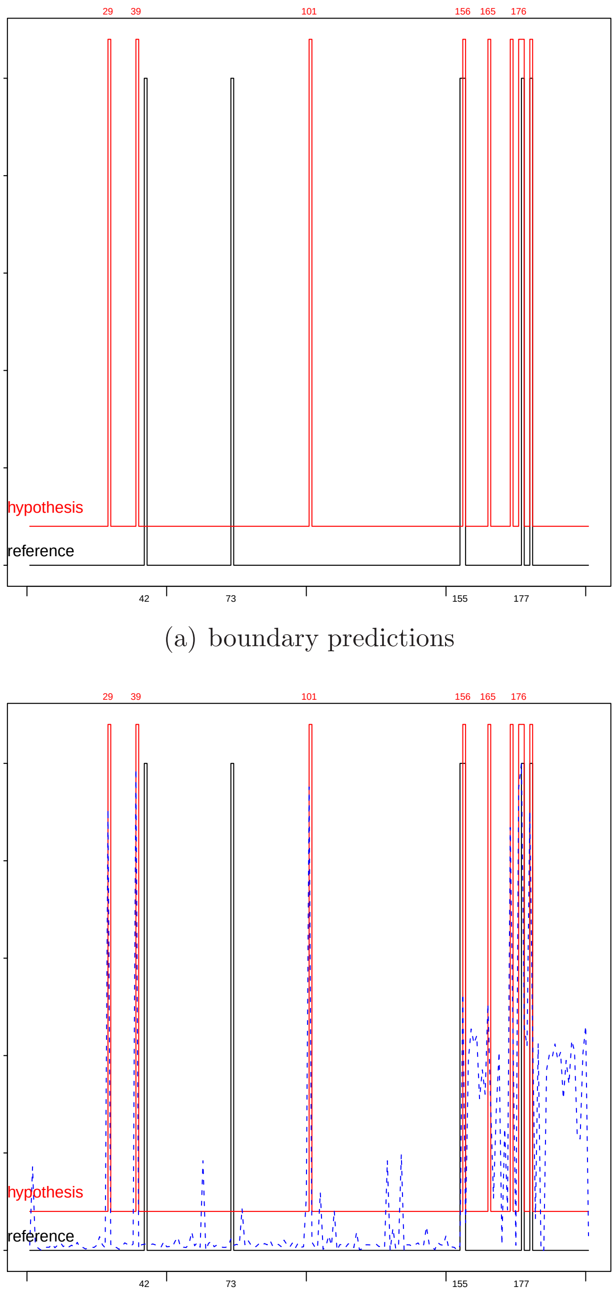
Feature Set	P_k	WD	ω
EP	0.408	0.54	1.78
EP_{VOC}	**0.365**	**0.496**	1.51
EP_{VOCP}	0.405	0.558	1.92
EP_{GAP}	0.408	0.599	2.48
Baseline$*$	0.506	0.565	1

$*$ accuracy of proportional uniform boundaries

Comparing Table 6.2 and Table 6.1 I see that C4.5 outperforms MAP NB with nearly each feature set on P_k and WD[7], but poor ω values of C4.5 undermines its performance. However, the best score of MAP NB comes with EP_{VOC} features, where $P_k = 0.365$ is much better than that of random baseline. EP_{VOC} predicts 1.5 times of the number of real boundaries ($\omega = 1.51$), so it contains 50% redundant positive predictions. This redundancy can be explained as false positive predictions from MAP NB classifier, as shown in Table 6.7(b) (Page 82).

Figure 6.2(a) shows a part of topic boundary prediction sequence from MAP NB. This plot has two advantages over C4.5 predictions in Figure 6.1. First, the number of positive predictions highly increases. Second, most predictions are located near to real boundaries. Both of the observations indicate that NB

[7] Segmentation is more accurate when P_k and WD approach 0.



(a) boundary predictions

(b) boundary predictions with probability plot in blue dash line

Figure 6.2: MAP Naïve Bayes prediction with *EP* features. (a) and (b) are from the same data set. In (b), blue dash line indicates posterior probability of predictions in positive class.

posterior probability density actively matches real boundary distribution.

Since the MAP NB classifier inevitably predicts redundant boundaries adjacent to each other, I probe NB posterior probability density for a solution. In Figure 6.2(b)[8], the black line indicates the positions of manually labeled "reference" topic boundaries, the red line is MAP NB predicted topic boundaries, and the blue dash line indicates posterior probability of positive class. In this figure, any instance with higher than 50% posterior probability of positive class is predicted as a boundary (red bars). MAP NB yields probability values with low precision, because of its impractical assumptions. But in most cases NB classifier assigns higher probability to the right class. This is why most non-boundary instances are correctly classified. In case I neither modify feature densities nor adjust NB assumptions, there are still options to reduce false positive predictions:

1. To merge closely adjacent predicted boundaries.

2. To modify MAP and increase the threshold of positive class predictions

Two closely adjacent boundaries (positive predictions) are obviously redundant, but classification may produce such output. It is necessary to filter off false positive boundaries by choosing one out of several consecutive positive predictions.

Besides adjacent false positive predictions, there are stand alone FP cases, which could not be filtered off by checking the predictions of nearby instances. However, probability prediction from MAP NB model could be used to trace positive predictions with relative low probability, which could be FP or TP. I filter off such cases and test if segmentation accuracy can be improved. For example, in Figure 6.2(b), instance No.165[9] is not a boundary instance by reference, but NB model generates incorrect probability (higher than 50%), then the prediction is wrong. If the probability threshold is increased to 80%, the false positive prediction on No.165 is avoided. In next section I introduce how to adapt probability thresholds with a view to improving performance.

[8] Data samples used in this figure is different from those in Figure 6.1, but (a) and (b) are from the same data samples.

[9] Vocalisation Event is the only unit in segmentation, instead of time and frame. No.76 refers to the 76th VE in the corresponding meeting record.

6.2.4 Thresholding for Naïve Bayes

MAP Naïve Bayes classifier is commonly observed to assign the right class to most instances, although its probability estimation is not precise. From MAP NB posterior probabilities shown in Figure 6.2(b), within the majority of correctly predicted boundaries, the probability values are approaching 1. So I make an assumption *MAP NB predictions with high probability values are mostly correct, but the ones with relatively low probability are less trustworthy.* Consequently, a collection of NB predictions with highest probabilities can be selected from MAP NB predictions, and reduce false positive cases. Two selection criteria are proposed here:

1. Proportional Threshold (**PT**) NB: only the top $n\%$ instances with highest probability are predicted as positive class, where n is the ratio of positive instances in training set.

2. Fixed Threshold (**FT**) NB: only the instances with probability higher than p are predicted as positive class. I set $p = 99\%$ in AMI.

Table 6.3: Classification accuracy of Fixed Threshold and Proportional Threshold Naïve Bayes classifier (with $NoAdj$ filtering). Each classifier is tested on 4 empty pause based feature sets.

		P_k	WD	ω
FT-NB	EP	0.341	0.41	0.553
	EP_{VOC}	0.328	**0.406**	0.654
	EP_{VOCP}	0.34	0.423	0.716
	EP_{GAP}	**0.326**	0.44	1.144
PT-NB	EP	0.378	0.471	0.962
	EP_{VOC}	0.355	0.434	0.755
	EP_{VOCP}	0.365	0.446	0.789
	EP_{GAP}	**0.344**	**0.429**	0.828
∗ Base		0.506	0.565	1

∗ Refer to baseline of proportional uniform boundaries.

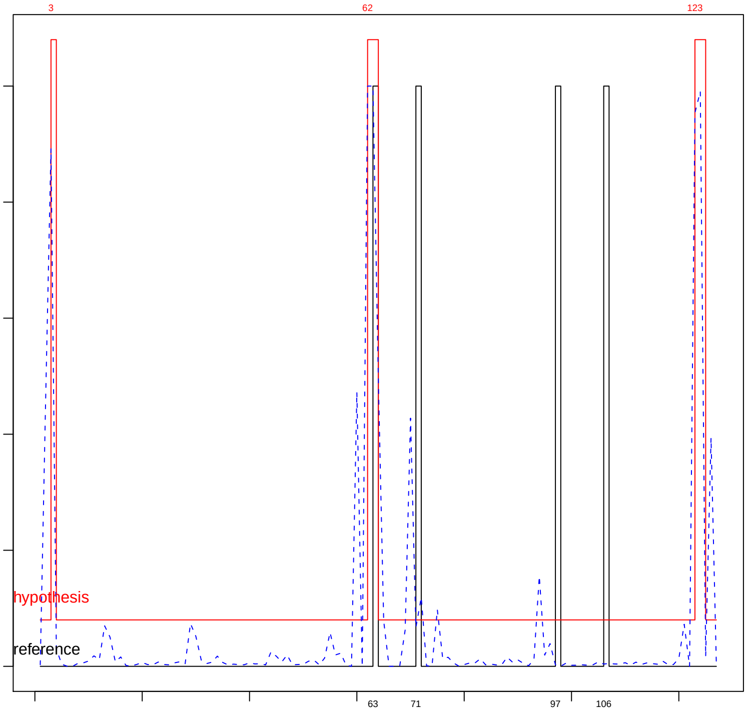
Applying PT and FT thresholds to EPs features (Equation 3.5) to (3.8), I obtain segmentation accuracy values in Table 6.3. Comparing Table 6.2 and Table 6.3, I have the following observations:

1. On each feature set, FT and PT NB generate higher segmentation accuracy and better ω than MAP NB.
2. FT and PT NB significantly reduce ω value and reduce false positive predictions.
3. The highest accuracy of thresholded classifiers is always obtained from the EP_{GAP} and EP_{VOC} feature sets, but best ω emerges with EP.
4. FT generates higher segmentation accuracy (lower P_k and WD) than PT in most cases.
5. PT generates ω closer to 1 than FT in most cases, so PT delivers the most similar quantity of boundaries as reference.

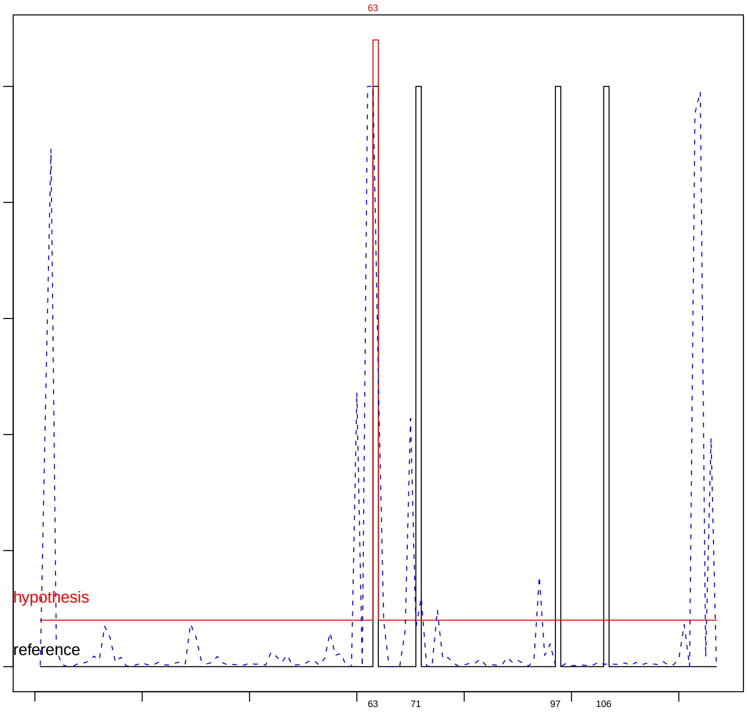
These observations agree with Domingos and Pazzani [1996] and confirm that segmentation improves when thresholds values are allowed to vary in NB. From Table 6.7(b) and Table 6.7(d), it is clear that PT and FT significantly decrease *false positive* predictions with respect to MAP NB. Although FT has higher segmentation accuracy, I recommend the use of PT instead. The reason is that PT predicts similar numbers of FP and FN, and the generated number of topic segments is similar to reference, which is better than predicting more segments than reality.

Figure 6.3(a) and Figure 6.3(b) show classification output from the MAP NB algorithm and PT NB algorithm separately with same instances. In Figure 6.3(a), the vocalisation instances No.3 and 123 are false positive predictions, but MAP cannot reject them with probability higher than 50%. In Figure 6.3(b) these two instances are rejected, because the ratio of positive instances in the training set is calculated, and in test set only the instances with top probability can be selected. Proportional thresholding effectively reduces false positive predictions. On the other hand, the true positive instances 71, 97, 106 are neglected in both algorithms. These false negative instances have relatively low probability, and the error can not be corrected by modified threshold classifiers.

As a conclusion, probability based Naïve Bayes classifier not only outperforms MAP NB, but also owns advantage over decision tree classifiers (Section



(a) MAP Naïve Bayes output



(b) PT Naïve Bayes output

Figure 6.3: Compare MAP and PT Naïve Bayes predictions with EP features, the blue dash line is probability plot (In this figure, instances are different with previous figures, in order to highlight FP predictions, but (a) and (b) are from the same instances and the comparison is valid.)

6.2.2). Naïve Bayes classifier has been validated for its potential of improving segmentation accuracy with threshold modifications.

6.2.5 CRF

Conditional random fields (CRF) are introduced to topic segmentation with two considerations (Section 4.2.3). First, CRF is a probabilistic model which is more appropriate than rule based models in segmentation experiments (e.g., naïve Bayes and decision tree). Second, CRF naturally accounts for sequential information among samples. According to these advantages, I test CRF model on the AMI corpus. Since CRF requires categorical features for feature function (Equation 4.11), continuous variables are categorised at first (Section 6.2.5.1).

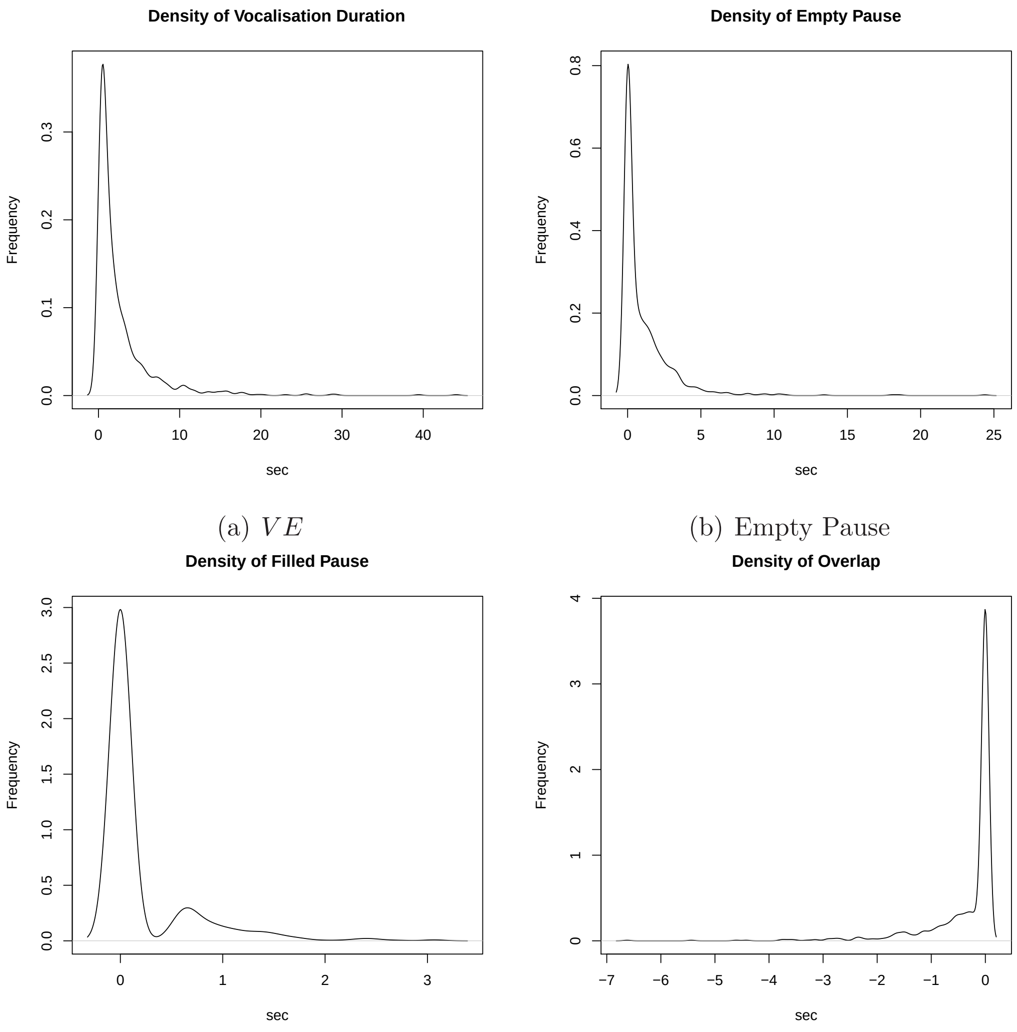
6.2.5.1 Categorised conversational features

Conversational features discussed in Section 3.1.2 are mostly continuous variables, except speaker ID and file names. Since CRF is a linear classifier, continuous features need to be normalised. I check normality of numerical features on meeting ES2002d[11] from AMI corpus. High skewness[12] of 4 features is observed in Table 6.4 as well as Figure 6.4. Normality can be improved with logarithmic transformation. However, there are still many extreme values (outliers) in log transformed variables. Therefore I categorise d, d_e, p_f p_e and o with predefined levels to minimise the influence of extreme values. The categorisation levels are recorded in Table 6.5. In classification experiment, the CRF model uses all the categorised features, which are introduced in Section 3.1.2, except t.

Categorisation is conducted with arbitrary levels for each continuous variable, where category levels are short at high density intervals. Table 6.5 shows category definition on 4 numerical features and the label of each level. In the categorised FP_{VOCP} feature set (Equation 3.3), VE_c start time t is excluded and the class label is converted from 0/1 to yes/no. Since the distribution of d, p_e, p_f and o are different, it is preferable to specify different categories for each variable.

[11] ES2002d is a typical meeting in the AMI corpus and the choice of meeting is with no preference.

[12] For Normal distribution, $Skewness = 0$



(a) VE

(b) Empty Pause

(c) Filled Pause

(d) Overlap

Figure 6.4: Density probability distribution of (a)Vocalisation Event duration, (b) Empty pause duration, (c) Filled pause duration and (d) Overlap duration in the unit of Second. Each distribution is skewed and need to be normalised.

Table 6.4: Simple statistics on four features of meeting ES2002d: VE duration d, empty pauses duration p_e, filled pause duration p_f, and overlap duration o (unit is Second)

	Mean	Median	Min	Max	Skewness
d	2.65	1.21	0.03	44.06	4.18
p_e	1.08	0.21	0	24.38	4.78
p_f	0.2	0	0	3.07	2.75
o	-0.34	0	-6.62	0	-3.46

Table 6.5: Categorisation of conversational features: (a) d, (b) p_e, (c) p_f and (d) o. In each sub-table, the left column records the interval of a continuous variable, and the right column records the category level assigned to that interval. As a consequence, continuous variables d, p_e, p_f and o are transformed to discrete variables and can be used for CRF training and testing.

(a)

d (sec)	Levels
$0 \sim 1$	VOC0
$1 \sim 3$	VOC1
$3 \sim 5$	VOC3
$5 \sim 7$	VOC5
$7 \sim 10$	VOC7
$10 \sim 15$	VOC10
$15 \sim 20$	VOC15
≥ 20	VOC20

(b)

p_e (sec)	Levels
0	NoEP
$0 \sim 1$	EP0
$1 \sim 3$	EP1
$3 \sim 5$	EP3
$5 \sim 7$	EP5
$7 \sim 10$	EP7
$10 \sim 15$	EP10
$15 \sim 20$	EP15
≥ 20	EP20

(c)

p_f (sec)	Levels
0	NoFP
$0 \sim 1$	FP0
$1 \sim 2$	FP1
$2 \sim 3$	FP2
$3 \sim 5$	FP3
$5 \sim 7$	FP5
$7 \sim 10$	FP7
≥ 10	FP10

(d)

o (sec)	Levels
0	NoOverlap
$-1 \sim 0$	Overlap0
$-2 \sim -1$	Overlap1
$-3 \sim -2$	Overlap2
$-5 \sim -3$	Overlap3
$-10 \sim -5$	Overlap5
≤ -10	Overlap10

6.2.5.2 CRF and Naïve Bayes

I categorise 4 feature sets for testing the CRF model: EP (Equation 3.5), EP_{VOC} (Equation 3.6), EP_{VOCP} (Equation 3.7), EP_{GAP} (Equation 3.8). Then I compare CRF topic segmentation accuracy with PT NB classifier. In this experiment, 30 meetings are collected from all phases of AMI discussion as training and test sets. In order to enrich training instances, I perform N-fold (N=30) cross-validation on a concatenated list of instances from all 30 meetings, to generate the accuracy value for each feature set, instead of classifying instances from each single meeting and offer the mean value of accuracy.

Table 6.6: Segmentation accuracy from CRF and PT NB

		P_k	WD	ω
CRF	EP	**0.381**	**0.403**	0.073
	EP_{VOC}	0.391	0.409	0.031
	EP_{VOCP}	0.385	0.404	0.045
	EP_{GAP}	0.386	0.411	**0.102**
PT-NB	EP	0.378	0.471	0.962
	EP_{VOC}	0.355	0.434	0.755
	EP_{VOCP}	0.365	0.446	0.789
	EP_{GAP}	**0.344**	**0.429**	0.828

Comparing Table 6.6 and Table 6.3, I can make a few interesting observations:

1. CRF generates better WD values than PT NB on all feature sets. On P_k metric, PT NB is better.

2. CRF predictions have very low ω values.

3. The effect of horizon is highly influenced by classifier. For PT NB, *Vocalisation Horizon* and *GAP Horizon* both enhance their base feature sets. EP_{GAP} generates best P_k, which is also true on FT NB. But VH weakens CRF on EPs feature sets.

Although CRF seems to outperform NB with its WD values, a closer look at the confusion matrix (Table 6.7) reveals hidden problems. CRF has 39 positive predictions, in which 8 are correct. But the actual positive instances are 384. Unbalanced FP and FN predictions contradict the definition of *Goodness* proposed in Section 5.1. In a data set with 30 meetings and 384 segments, 39 indicated topic boundaries would provide little guidance to a meeting browser. CRF, as a classifier accommodating sample dependencies, has not met topic segmentation requirements so far. PT and FT NB are better choices.

Table 6.7: Confusion matrix of four classifiers: (a) CRF, (b) MAP NB, (c) Fixed Threshold NB (p=0.99) and (d) Proportional Threshold NB on EP_{GAP}. (a) CRF has the lowest number of true positive predictions.

(a)

a	b	← pred
8	376	a=1
31	12933	b=0

(b)

a	b	← pred
143	240	a=1
1262	11675	b=0

(c)

a	b	← pred
83	300	a=1
505	12432	b=0

(d)

a	b	← pred
59	324	a=1
330	12607	b=0

6.2.6 Ensemble classifiers

Ensemble classifiers are designed to learn more expressive concepts than a single classifier (Section 4.2.4). Since the highly unbalanced AMI data set challenged classical classifiers (e.g., decision tree), I am interested in assessing ensemble classifiers for topic segmentation. Both Bagging and Boosting rely on collective votes of base classifiers in multiple iterations, so the choice of base classifier inevitably influences classification results. Here C4.5 decision tree and MAP naïve Bayes are selected as base classifiers[12].

Since C4.5 suffers from the highly unbalanced AMI corpus (Section 6.2.2), here I test whether ensembles improve C4.5 and how much they do (Section

[12] PT NB and FT NB outperform MAP NB in many cases, and could be used as base classifier as well. I plan to implement updated algorithms in future work.

6.2.6.2). The complexity of the decision tree determines highly influences the tree's prediction power, so I select the optimal branching factor before running ensemble experiments (Section 6.2.6.1).

MAP naïve Bayes classifier exhibits topic boundary prediction power(Section 6.2.3), and improves segmentation accuracy with thresholding techniques(Section 6.2.4). I am interested in using MAP NB as a base classifier of ensembles, and compare its segmentation power with C4.5 base classifier. Moreover, it would be interesting to know whether ensembles improve MAP NB more than thresholding. The experiments are listed in Section 6.2.6.3.

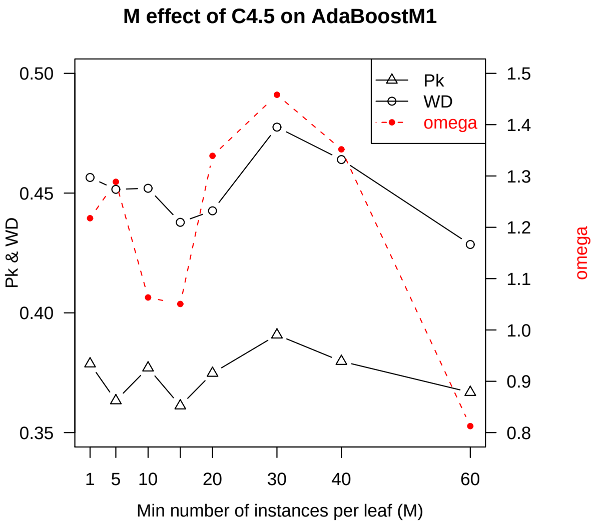
6.2.6.1 The effect of Min instances per leaf in C4.5

The prediction power of unpruned C4.5 decision trees varies with M (Min number of instances per leaf). For a small M (e.g., $M = 1$), the tree may be too complex and over-fit training set, so as to be less predictive on test set. I test M from 1 to 60[13] on AdaBoostM1 (Figure 6.5(a)) and Bagging 6.5(b)) separately. For simplicity, only EP features are involved. AdaBoostM1 reaches best P_k when $M = 60$, but ω drops to 0.8 and is expected to drop further along the trend. The overall best choice is $\boldsymbol{M}$**=15**, where ω is closest to 1 and P_k, WD are relatively small. Similarly, $\boldsymbol{M}$**=5** is best for Bagging. These two M values are selected for more complex experiments of C4.5 classifier.

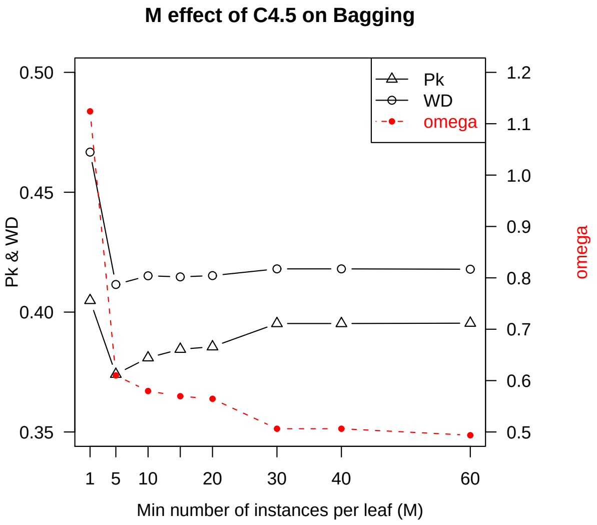
6.2.6.2 Using decision tree base classifier

In this section, ensemble classifiers are tested against 3 feature sets EP, EP_{VOC} and EP_{GAP} from 30 AMI meetings. AdaBoostM1 and Bagging take branching factor $\boldsymbol{M}$**=15** and $\boldsymbol{M}$**=5** each. Through these experiments I expect to check *Horizon* effect upon ensemble classifiers. As indicated in Section 6.2.3, two or more closely adjecent topic boundaries are redundant. **NoAdj** filtering algorithm only reserves the first boundary in a queue connected boundaries, and reduces

[13] M stands for the minimum number of instances per leaf and controls the complexity of decision tree. If M approaches 1, the tree may overfit training data and lacks generality. On the other hand, if M is too big (such as $M > 60$), an existing leaf must have at least 60 instances satisfying its rule, so that the tree may be too simple or only contain a stub. I assume $M > 60$ is not practical in this experiment.



(a) AdaBoostM1



(b) Bagging

Figure 6.5: Accuracy of ensemble classifiers on EP feature with various C4.5 leaf settings

false positive error. I apply this filtering algorithm with C4.5 and naïve Bayes based ensembles.

Table 6.8: Segmentation accuracy of unpruned C4.5 with $M = 15$

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.597	0.968	11.9	0.536	0.846	10.85
EP_{VOC}	0.607	0.987	25.75	0.485	0.72	25.27
EP_{GAP}	0.58	0.939	25.31	0.456	0.671	24.8

Table 6.9: Segmentation accuracy of AdaBoostM1 with C4.5 base classifier, $M = 15$

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.361	0.438	1.05	**0.361**	**0.433**	**0.71**
EP_{VOC}	0.391	0.434	0.41	0.391	0.434	0.39
EP_{GAP}	0.351	0.397	0.28	0.351	0.392	0.26

Table 6.10: Segmentation accuracy of Bagging with C4.5 base classifier, $M = 5$

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.374	0.412	0.61	0.374	0.412	0.241
EP_{VOC}	0.392	0.426	0.27	0.392	0.426	0.24
EP_{GAP}	0.385	0.407	0.065	0.385	0.407	0.057

Table 6.8 is a re-visit of C4.5 decision trees with branching factor M=15, which produces much worse accuracy than $M = 5$ in Table 6.1. Higher M score over prunes decision trees and makes it too simple. $M = 15$ leads to ω around 25, a prediction with too many false positives. On the contrary, Bagging (Table 6.10) and Boosting (Table 6.9) increase boundary prediction accuracy significantly. Especially with EP features, AdaBoostM1 generates ω=0.71, which makes an acceptable segmentation. Carefully examining Bagging and Boosting results, I have the following observations:

1. VOC_Horizon and GAP_Horizon have negative effect on the number of predicted boundaries. Each of them corresponds to $\omega < 0.5$, which seriously drops the value of boundary prediction for audience.

2. **NoAdj** plays a significant role on EP features, where ω drops 0.34 for AdaBoostM1 and 0.37 for Bagging after purification. But for Horizon features, **NoAdj** drops ω equal or less than 0.03.

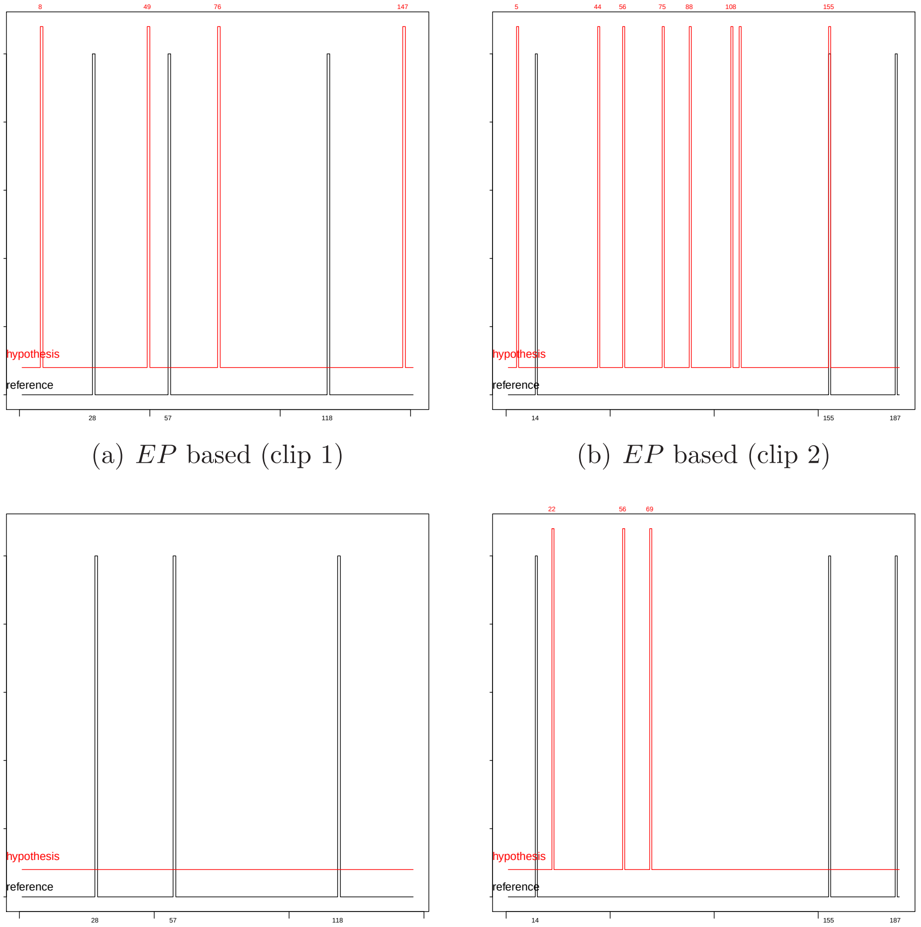
3. Purified Bagging predictions own $\omega < 0.25$ for each feature set, so Bagging with C4.5 is not helpful for boundary prediction.

The first observation suggests that ensemble classifiers with rule based base classifier are sensitive to the dimension of features, or they are sensitive to correlated features. Horizon feature conveys sequential information, but it harms C4.5 based ensembles.

As regards the second observation, Since the adjacent boundaries are almost all false positive predictions, and **NoAdj** significantly drops ω for Bagging and Boosting with EP feature, I expect an increase on P_k and WD. But the fact is P_k and WD are nearly the same for *Original* and *NoAdj*, which means a high portion of remaining boundaries are still misplaced.

In addition, one might ask if the minor effect of **NoAdj** on Horizon features mean that Horizon especially produces less false positives. The answer is negative, because Horizon does not significantly improve P_k and WD. An illustration of horizon effect in C4.5 based AdaBoostM1 is shown in Figure 6.6. From (a) to (c) VOC_Horizon removes all near-miss prediction and leaves no prediction, but from (b) to (d) VOC_Horizon diminishes many true redundant predictions. Horizon has both positive and negative effect, but in general it misses too many predictions (shown in ω).

The best combination from C4.5 based ensembles is Boosting with EP, where P_k=0.361, WD=0.433 and ω=0.71. In order to have a fair evaluation of Boosting over thresholded naïve Bayes, I choose the one with a similar ω value. EP_{VOCP} based FT-NB generates P_k=0.34, WD=0.423 and ω=0.716. So FT-NB reaches higher segmentation accuracy than C4.5 based Boosting, and consequently the latter classifier is not of first choice.



(a) EP based (clip 1)

(b) EP based (clip 2)

(c) EP_{VOC} based (clip 1)

(d) EP_{VOC} based (clip 2)

Figure 6.6: Compare EP and EP_{VOC} features for C4.5 based AdaBoostM1 (with **NoAdj**)

6.2.6.3 Using naïve Bayes base classifier

Another design of Ensemble classifier is to use MAP naïve Bayes as the base classifier, because NB has been proved of suitability on topic boundary detection. In this experiment, EP, EP_{VOC} and EP_{GAP} feature sets are tested and I get topic boundary predictions from MAP naïve Bayes (Table 6.11), AdaBoostM1 (Table 6.12) and Bagging (Table 6.13). I observe a few differences with respect to C4.5 based ensemble classifiers:

Table 6.11: Segmentation accuracy of MAP naïve Bayes classifier

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.408	0.545	2.04	0.408	0.54	1.78
EP_{VOC}	0.365	0.523	2.43	**0.365**	**0.496**	**1.5**
EP_{GAP}	0.408	0.622	3.66	0.408	0.599	2.47

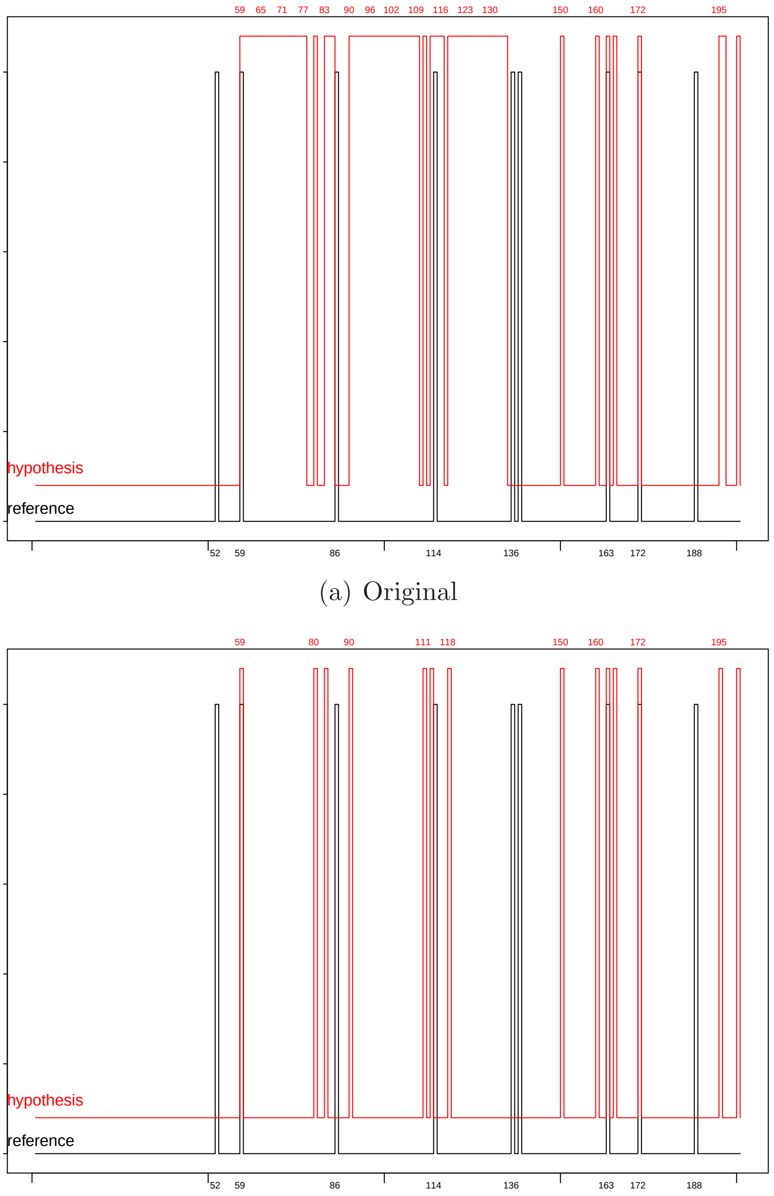
Table 6.12: Segmentation accuracy of AdaBoostM1 with MAP naïve Bayes base classifier

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.391	0.497	3.4	**0.385**	0.482	1.44
EP_{VOC}	0.406	0.503	3.7	0.387	**0.468**	**0.98**
EP_{GAP}	0.406	0.526	3.95	0.403	0.496	1.32

Table 6.13: Segmentation accuracy of Bagging with MAP naïve Bayes base classifier

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.404	0.538	2.03	0.404	0.533	1.73
EP_{VOC}	0.365	0.522	2.36	**0.365**	**0.495**	**1.48**
EP_{GAP}	0.4	0.617	3.67	0.4	0.592	2.47

1. **NoAdj** plays a significant role for each feature set and each ensemble classifier. Through **NoAdj**, ω drops at most 38%, 73% and 37% for MAP



(a) Original

(b) **NoAdj** filtered

Figure 6.7: Effect of **NoAdj** on NB based AdaBoostM1

NB, AdaBoostM1 and Bagging. Comparing with C4.5 based ensembles, NB based ensembles generate much more spurious boundaries.

2. VOC_Horizon improves segmentation accuracy but GAP_Horizon impairs it with both Bagging and Boosting.

3. EP_{VOC} performs better than other two feature sets with each classifier. With EP_{VOC}, Bagging offers minor improvement over MAP NB, but Boosting significantly improves WD and ω.

From these facts, I notice the necessity of **NoAdj** method over probability based classifiers, which is more prone to producing adjacent boundaries. The filtered boundary sequences in Figure 6.7(b) are mostly close to real boundaries, although there are still false positives. VOC_Horizon has a positive effect with naïve Bayes classifier (Section 6.2.3 and 6.2.4), and it also enhances NB based Ensemble classifier. As a conclusion, Boosting is superior over Bagging with MAP NB base classifier, and the latter has little advantage over its base classifier.

6.2.6.4 Conclusion

The segmentation effect of C4.5 based ensemble classifiers varies with the complexity of the base classifier. I tested with EP features, and set the minimum number of instances per leaf to be $M = 15$ for Boosting and $M = 5$ for Bagging as optimal choices. Experiments with other feature sets follow this setting.

With C4.5 base classifier, Vocalisation Horizon is redundant. Boosting has much better segmentation accuracy than C4.5 alone, and produces the best accuracy on EP features. But this accuracy is weaker than FT naïve Bayes algorithm. Bagging produces too low ω values and is therefore unsuitable for this task.

Naïve Bayes based ensemble classifiers have worse P_k and WD values than C4.5 based one, but NB always produces an ω score very close or higher than 1, which is better than under estimates. Boosting has better WD and ω level than NB alone, but Bagging has equivalent performance as NB alone. Comparing NB based Boosting with PT-NB, Boosting has advantage on ω but is weaker on WD and P_k. On the other hand, NB based ensembles tend to produce adjacent boundaries. The **NoAdj** algorithm effectively reduces such false positive predictions.

Generally, ensemble classifier shows advantages over its base classifier used alone, and naïve Bayes based ensembles are of practical use. It has comparable performance as thresholding naïve Bayes classifiers.

6.3 Feature selection for topic segmentation

In Section 6.2 a comprehensive study of classification approaches is conducted for content-free topic boundary detection. The study focuses on the comparison of classification algorithms with respect to various feature sets. However, I confine the study with four feature sets (Equation (3.5) to (3.8)) for convenience of analysis. In this section, I probe more vocalisational and acoustic features as well as their *Vocalisation Horizon* form, in order to evaluate their function with topic segmentation.

6.3.1 Empty pause and filled pause

In order to scrutinize the effect of *pause* on topic boundary locations, I distinguish pauses as *empty pause* and *filled pause* (introduced in Section 3.1.2). The two types of pauses are not simply additive features on VE_c, because they modify the definition of a VE_c. If "en" is treated as a filled pause and is extracted from a piece of continuous talk, this continuous vocalisation is separated into two new vocalisations. On the other hand, if I only identify empty pauses, the separation does not happen.

In this section, I examine these two settings through FT and PT naïve Bayes classifiers and evaluate whether *filled pause* has a positive effect on segmentation. Table 6.14 and 6.15 show segmentation results from VE_c with filled pauses and with only empty pauses, with respect to various feature settings. **NoAdj** method has been applied.

Two facts are discernable here. First, for each VE_c feature set (i.e., VOC, GAP) *EPs* has higher segmentation accuracy than *FPs* with respect to both P_k and WD. Moreover, *EPs* have advantage in most ω scores.

This result shows that filled pauses are not particularly effective as classification features, which contradicts claims by Swerts et al. [1996] on the utility of filled pauses. Empty pauses produce better results than filled pauses. We can also say that reserving long VE_c instead of breaking them at filled pause yields higher segmentation accuracy.

Second, the GAP features are superior to VOC with any metric in PT NB, and GAP are partially superior in FT NB, on P_k and ω. So, with Bayesian classifier, the horizon of empty pause and overlap are more predicative than vocalisation horizon, no matter if filled pause is in consideration. But with Ensemble classifiers, VOC produces better results (see Table 6.9 and 6.10, Table 6.12 and 6.13). A possible explanation is that GAP features have 6 more dimensions than VOC, and the dimensions are not independent. Higher complexity of feature space adversely influences most classifiers, but less so on naïve Bayes models.

Table 6.14: Classification accuracy of Fixed Threshold (FT) Naïve Bayes classifier on EPs and FPs features (with $NoAdj$)

	P_k	WD	ω
EP	0.341	0.41	0.553
EP_{VOC}	0.328	**0.406**	0.654
EP_{VOCP}	0.34	0.423	0.716
EP_{GAP}	**0.326**	0.44	1.144
FP	0.386	0.433	0.455
FP_{VOC}	0.383	0.422	0.298
FP_{VOCP}	0.391	0.455	0.501
FP_{GAP}	0.366	0.466	**1.034**

6.3.2 Augmented Horizon features

Vocalisation horizon and GAP horizon are tested upon with various classification schemes. Results demonstrate that horizon has substantial effect of improving segmentation accuracy. In previous experiments, *Vocalisation horizon* only utilises the basic property: *duration* of VE. Now I expect to use more VE features in study, and thoroughly analyse the potential of horizon. Speaker role is selected as alternatives in this section.

Table 6.15: Classification accuracy of Proportional Threshold (PT) Naïve Bayes classifier on EPs and FPs features (with $NoAdj$)

	P_k	WD	ω
EP	0.378	0.471	**0.962**
EP_{VOC}	0.355	0.434	0.755
EP_{VOCP}	0.365	0.446	0.789
EP_{GAP}	**0.344**	**0.429**	0.828
FP	0.42	0.499	0.954
FP_{VOC}	0.418	0.484	0.797
FP_{VOCP}	0.411	0.484	0.861
FP_{GAP}	0.395	0.471	0.804

6.3.2.1 Speaker role

Briefly speaking, *speaker role horizon* integrates the speaker role of previous or following VEs as a feature of current VE. Since FPs based features are not necessary, I only expand EPs features. They are formally represented from EP_{ROLE} (Equation 6.1) to EP_{ROLE_GAP} (Equation 6.4).

$$\begin{aligned} EP_{ROLE} &= (s, t, d_e, s_{-n}, ..., s_{-1}, s_1, ..., s_n) && (6.1)\\ EP_{ROLEP} &= (s, t, d_e, p_e, o, s_{-n}, ..., s_{-1}, s_1, ..., s_n) && (6.2)\\ EP_{ROLEP\_VOC} &= (s, t, d_e, p_e, o, d_{-n}, ..., d_{-1}, d_1, ..., d_n, \\ &\qquad s_{-n}, ..., s_{-1}, s_1, ..., s_n) && (6.3)\\ EP_{ROLE\_GAP} &= (s, t, d_e, p_e, o, u_{-n}, ..., u_{-1}, u_1, ..., u_n, \\ &\qquad s_{-n}, ..., s_{-1}, s_1, ..., s_n) && (6.4) \end{aligned}$$

Equations (6.1) to (6.4) show vocalisation features without filled pauses, where $u_i = (p_{ei}, o_i)$. I name these features together as EP_{ROLE} based features. In all these equations, s is the identifier for one speaker role, s_i is speaker role of the i^{th} VE_c preceding (i<0) or following (i>0) VE_c, t is the start time of current VE_c, d is its duration (d_e refers to VE_c duration without filled pause), p_e is duration of empty pause, o is the negative value of overlap duration, d_i is the duration of the i^{th} VE_c preceding (i<0) or following (i>0) VE_c, u_i refers to empty pause and overlap preceding or following VE_c (for VE_c without filled pause) and I set $n = 3$ as the length of the context (or "horizon") spanned by the VE.

Table 6.16: Segmentation accuracy of PT NB with ROLE Horizon related features

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.378	0.473	0.99	0.378	0.471	0.962
EP_{ROLE}	0.379	0.47	1.01	0.379	0.467	0.974
EP_{ROLEP}	0.388	0.478	1.01	0.388	0.478	0.98
EP_{ROLEP_VOC}	0.354	0.448	1.02	0.354	0.431	0.763
EP_{ROLE_GAP}	0.376	0.469	1.01	0.374	0.45	0.75

Table 6.17: Segmentation accuracy of MAP NB based AdaBoostM1 with ROLE Horizon related features

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.391	0.497	3.4	0.385	0.482	1.44
EP_{ROLE}	0.386	0.476	1.22	0.386	**0.471**	**1.1**
EP_{ROLEP}	0.413	0.505	1.96	0.413	0.502	1.34
EP_{ROLEP_VOC}	0.386	0.502	2.57	**0.38**	0.478	1.11
EP_{ROLE_GAP}	0.419	0.557	3.69	0.419	0.541	1.72

In Section 6.2 I found that thresholding naïve Bayes classifier and naïve Bayes based Boosting classifier are mostly successful on topic segmentation. So I test *ROLE* Horizon effect with these two classifiers. From all PT NB results (Table 6.16), EP is a benchmark, which yields nearly perfect ω and moderate P_k, WD. Comparing with EP, EP_{ROLE} has limited effect on each metric. However, other feature sets drop either ω or P_k, WD. With the increasing model complexity, ω decreases. When I have cross comparison against VOC/GAP Horizon in Table 6.3, it is also hard to conclude that *ROLE* Horizon offers advantage. It is fair to conclude that *ROLE* is an useful alternative of VOC with PT NB, since *ROLE* predicts a better ω.

On Boosting algorithm (Table 6.17), EP_{ROLE} and EP_{ROLEP_VOC} predict better than EP, but the results are still worse than EP_{VOC} in Table 6.12, which obtains $\omega = 0.98$. Other combinations of $ROLE$ Horizon are even worse. Then VOC Horizon fits Boosting better than $ROLE$ Horizon.

6.3.2.2 Acoustic features

As suggested in Section 3.1.2.3, pitch and intensity features may signal segment boundaries. Here I test the importance of pitch in relation to topic boundary identification. In order to accommodate pitch information as a VE feature, I use the mean value of pitch during a VE, instead of extracting patterns of pitch variation as [Shriberg et al., 2000]. I expect to follow Shriberg's methods in future research.

$$\begin{aligned}
EP_{PP} &= (s, t, d_e, p_e, o, pch) && (6.5)\\
EP_{PITCH\_P} &= (s, t, d_e, p_e, o, pch, pch_{-n}, ..., pch_{-1}, pch_1, ..., pch_n) && (6.6)\\
EP_{VOC\_PP} &= (s, t, d_e, p_e, o, pch, d_{-n}, ..., d_{-1}, d_1, ..., d_n) && (6.7)\\
EP_{VOC\_PITCH} &= (s, t, d_e, p_e, o, pch, d_{-n}, ..., d_{-1}, d_1, ..., d_n,\\
&\qquad pch_{-n}, ..., pch_{-1}, pch_1, ..., pch_n) && (6.8)
\end{aligned}$$

Equations (6.5) to (6.8) show pitch horizon related features without filled pauses. I name these features together as EP_{PITCH}. Most of the symbols in equations are same as those in Section 6.3.2.1, except pch and pch_i which mean pitch mean value in a VE and pitch horizon in adjacent VE. I set $n = 3$ as the length of the context (or "horizon") spanned by the VE.

Table 6.18: Segmentation accuracy of PT NB with PITCH Horizon related features

	Original			NoAdj		
	P_k	WD	ω	P_k	WD	ω
EP	0.378	0.473	0.99	0.378	0.471	0.962
EP_{PP}	0.402	0.501	1.01	0.402	0.501	0.99
EP_{PITCH_P}	0.408	0.491	1.0	0.408	0.486	0.94
EP_{VOC_PP}	0.378	0.475	1.0	0.378	0.475	0.996
EP_{VOC_PITCH}	0.382	0.475	1.0	0.382	0.475	0.996

In experiment, VE separated pitch values are extracted from 23 AMI meetings recordings. Based on previous cases, PT NB is selected as the first classifier to test pitch horizon effect. Results in Table 6.18 are not pleasant, since they present very similar or even worse accuracy than the basic EP feature. Then it is not necessary to extract pitch feature and compose the complicated horizon features.

It is not responsible to simply discard all acoustic features for content-free topic segmentation, I expect to modify the approach of using pitch, and include more features such as voice intensity and prosody.

6.4 Additional experiments

I have in-depth study on the AMI corpus for the purpose of topic segmentation. As part of the research, the corpus inspires additional experiments on influential factors as well as supplemental objectives. A complete set of AMI meetings is always organised in 4 meetings phases. Since the objectives of each phase are predefined, I expect similarities of meeting content and structure among meetings of the same phase. Furthermore, such similarity among meetings is supposed to increase segmentation accuracy. I would like to verify such influence by separating meeting data set and implementing segmentation algorithms on meetings from the same phase (Section 6.4.1). If more homogeneous data set do improve segmentation accuracy, the relation between meeting structure and vocalisation event characteristics is supported.

6.4.1 The effect of meeting phases

I propose to study the relation between the homogeneity of meeting content and topic boundary detection. A prominent label of AMI meeting content is project phase. As introduced in Section 3.1, four phases of a design procedure are discussed in AMI meetings, and each meeting is related to only one phase. The objectives of phases are project introduction, functional design, conceptual design and detailed design. The predefined phases are named A, B, C, D. Since the objective of meetings is the same within each phase, I assume that meeting content and structure have more similarities within phases, and have more differences between phases. Furthermore, I hypothesise that such similarities will be useful to content-free topic boundary detection.

I take 12 meetings with EP_{VOC} features from each phase, and run 12-fold cross validation on a PT NB classifier. Accuracy from a mixture set of all phases (EP_{VOC} in Table 6.6) is used for comparison. Table 6.19 shows that P_k and WD from most single phases are higher than those in the mixture set. What's more, ω is higher in single phases, so the number of predictions in single sets is closer to reference. It is worth noting the segmentation accuracy increase introduced by meeting content group. Phase A and D meetings, corresponding to the project introduction and detailed design phases, generate the best scores. For certain groups of meetings, meeting content has significant effect on segmentation accuracy, when segmentation is based only on conversational features. This result shed some light on the good results obtained by Luz [2009] for a highly homogenous set of medical meetings.

Table 6.19: Accuracy of four types of meetings, from EP_{VOC} feature set and with PT NB

Meeting Type	P_k	WD	ω
A	**0.326**	0.430	0.891
B	0.331	0.427	0.896
C	0.350	0.448	0.865
D	0.328	**0.412**	0.828
All Meetings	0.355	0.434	0.755
Random Baseline	0.468	0.532	

6.5 Conclusion

In this chapter, I run experiments of content-free topic segmentation on the AMI corpus. The tasks are in three tracks, namely segmentation metric modification, classifier comparison and feature set selection. In the first track, I find deficiency of P_k and WD, both of which overlook the influence of boundary quantity toward segmentation accuracy. I supplement a novel metric $balancefactor\omega$ and it is tested to improves the fairness of segmentation metrics.

In the second track, various classification schemes (decision tree, naïve Bayes classifiers, conditional random fields and ensemble classifiers) are employed to distinguish topic boundary VE_s from other VE_s, so as to achieve topic segmentation. Unpruned C4.5 decision trees predict more than two times of the true boundaries, together with P_k, WD scores worse than random baseline. Then decision trees do not fit for topic segmentation. In the AMI corpus, positive instances only stand for 2.6% of population. Such unbalanced data set may be difficult for decision tree to generate robust rules.

MAP naïve Bayes classifier performs best with vocalisation horizon features. Its $\omega = 1.51$ and P_k, WD are better than baseline. NB is superior over decision trees, because NB classifier has the advantage of predicting higher probability estimation for the correct class, although the estimation may be not accurate [Domingos and Pazzani, 1996]. So, NB is more robust with unbalanced data set.

PT NB and FT NB are two modified versions of MAP NB (Section 6.2.4).PT NB filters boundary predictions and only keeps the n% instances with top probability as positive, where n is the ratio of positive instances in training set. FT NB applies 99% probability threshold other than 50% of MAP NB. Both versions exclude more false positive predictions than MAP NB and result in a ω closer to 1. Modifications of MAP NB are tested to be successful for AMI topic segmentation. Generally, FT NB has better P_k and WD values, but PT NB has better ω.

CRF predictions have very low ω values, hence it is not used as a topic segmentation method. On the other hand, ensemble classifiers exhibit higher segmentation accuracy than corresponding C4.5 and MAP naïve Bayes base classifiers. Comparing all tested classifiers, PT NB, FT NB and Boosting with MAP NB base classifier generate best P_k, WD scores with $\omega \approx 1$.

The last track of this chapter is to select feature sets for classification. Vocalisation horizon significantly improves segmentation accuracy with most classifiers tested, which verify the necessity of including features from neighboring instances in classification. I further test the horizon effect of other features. The effect of empty pause and overlap horizon (GAP horizon) is dependent on classifier selection, but empty pauses are always better than filled pauses. Speaker role horizon does not show deterministic advantage over VOC horizon, but it is an useful alternative for its better ω score.

From a number of experiments on the AMI corpus, I clarify the feasibility of applying classification schemes for content-free topic segmentation, together with a set of effective features. A good combination of classifier and features performs much better than others, although I cannot explore every combinations of features as well as classifiers. So far I do not have a global optimum solution. In this study, a very important finding is a complete set of segmentation metrics, which guarantee fair judgement of segmentation fitness. All of these indicate a systematic solution of content-free topic segmentation.

Chapter 7

MDTMs Experiments

A series of topic segmentation algorithms are carried out with the AMI corpus (Chapter 6), they validate the vocalisation structure as classification instances, and classification approaches over segmentation tasks.

In this chapter I conduct experiments with a corpus of MDTMs, which are sampled from real medical team meetings rather than simulated meetings, as in the AMI corpus. Similar settings from AMI experiments are employed in the MDTM experiments. I am interested whether there are differences in segmentation output, in order to evaluate whether the successful methods from simulated meetings are also effective on real data.

This chapter starts with an Exploratory Data Analysis (Section 7.1), where influential factors are identified. Statistical models (Section 7.2) indicate probability relations between topic boundary and key features. Then classification methods (Section 7.3) are investigated. These methods include naïve Bayes related models (Section 7.4.1) and ensemble classifiers (Section 7.4.2). In classification models, vocalisation horizon and evaluation metrics are under careful investigation.

7.1 Exploratory Data Analysis

Before constructing statistical models, I carry out exploratory data analysis (EDA). EDA helps to visualise feature properties and indicates the potentials of model

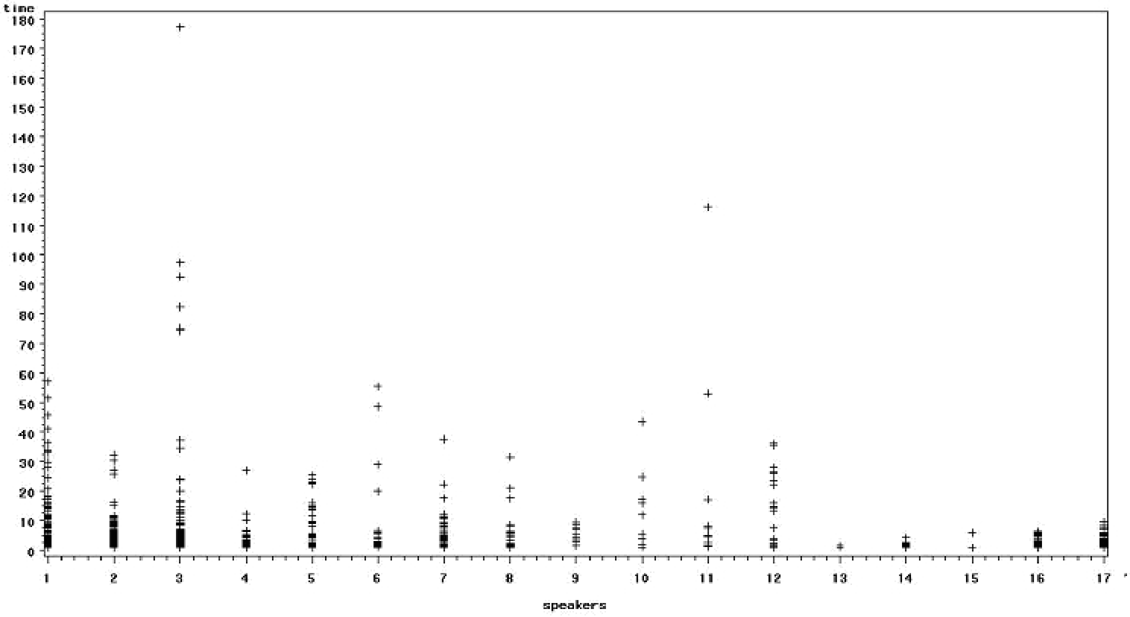


Figure 7.1: Vocalisation duration by each speaker

selection.

	case start =0	case start =1
Mean	6.77	15.51
Median	3.25	6.33
Std. Dev.	12.42	17.51
Min. value	0.82	0.98
Max. value	177.24	57.18
No. of observations	711	22

Table 7.1: Descriptive statistics of continuous variable: **vocalisation duration**

From Table 7.1, I see the mean vocalisation duration has clear difference between the vocalisations leading a case discussion and those not. This fact indicates *vocalisation duration* can be a useful predictor in regression model.

Figure 7.2 shows that a large portion of vocalisations range from 0 to 10 seconds, but their distribution is highly right-skewed with several extreme values. These extreme values weaken normality of the overall distribution, in both MDTM and AMI corpus. A common way to correct their influence is *log* transformation.

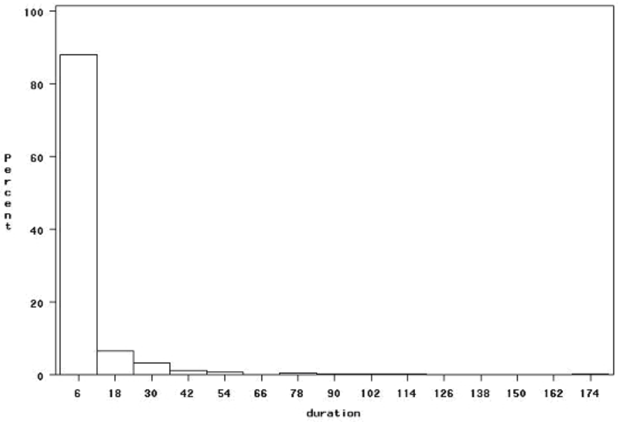


Figure 7.2: Vocalisation duration by percentile

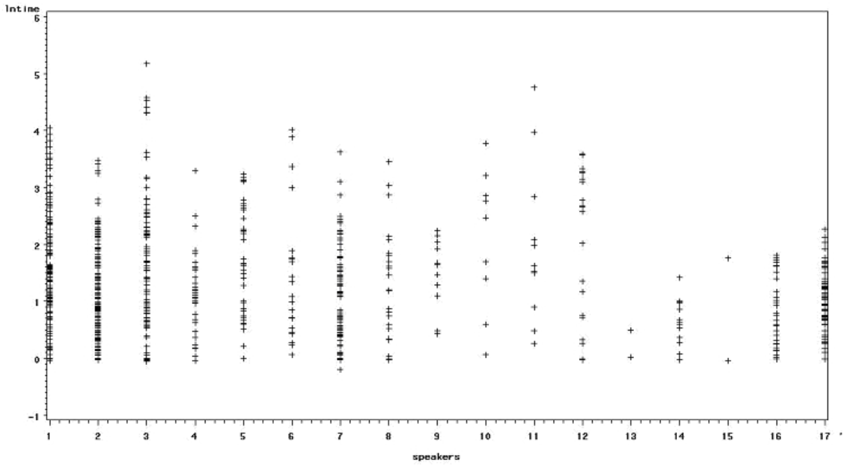


Figure 7.3: Log transformed vocalisation duration by each speaker

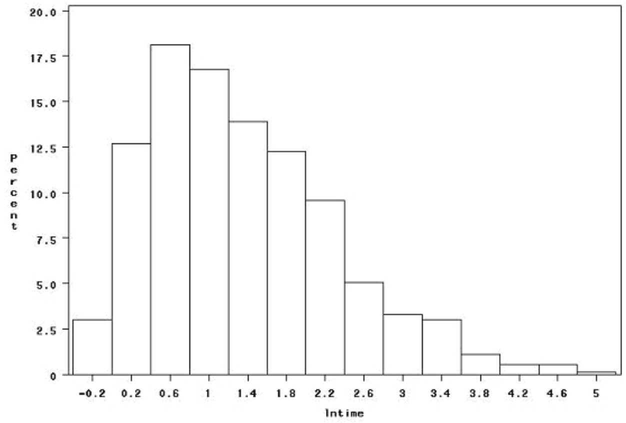


Figure 7.4: Log transformed vocalisation duration by percentile

Comparing Figure 7.2 and Figure 7.4, I see that *log* transformation improves the normality of vocalisation duration in a large scale, although *log* transformed vocalisation duration does not satisfy normality (Shapiro-Wilk test [Shapiro and Wilk, 1965] has $W = 0.947$ with $p < 0.05$) either. The choice of vocalisation duration or its log form is determined by statistical models in use. A multiple linear regression model requires that each numerical variable follows normal distribution, so that the residuals (model predicted values minus observations) are distributed normally. But Multiple logistic regression does not demand normality from variables, where the response variable is binary.

Each vocalisation event (VE) has a speaker ID related to it, the categorical variable speaker ID can be one predictor for statistical models. However, since MDTMs are highly structured meetings, they are organised based on meeting participant roles. For example, the respiratory multidisciplinary team is made up of three respiratory medical and two thoracic surgical teams, oncologist, radiologists, pathologists, radiation oncologist, nurse specialists, physiotherapist, radiation therapist(s), database managers and technical assistant [Kane and Luz, 2006]. In case someone is absent of a meeting, another person with the same speciality will take the role. In different patient case discussions (PCDs), the participant of the same role may differ. In other words, one pathologist only look after designated patients. With this background, I analyse MDTMs based on

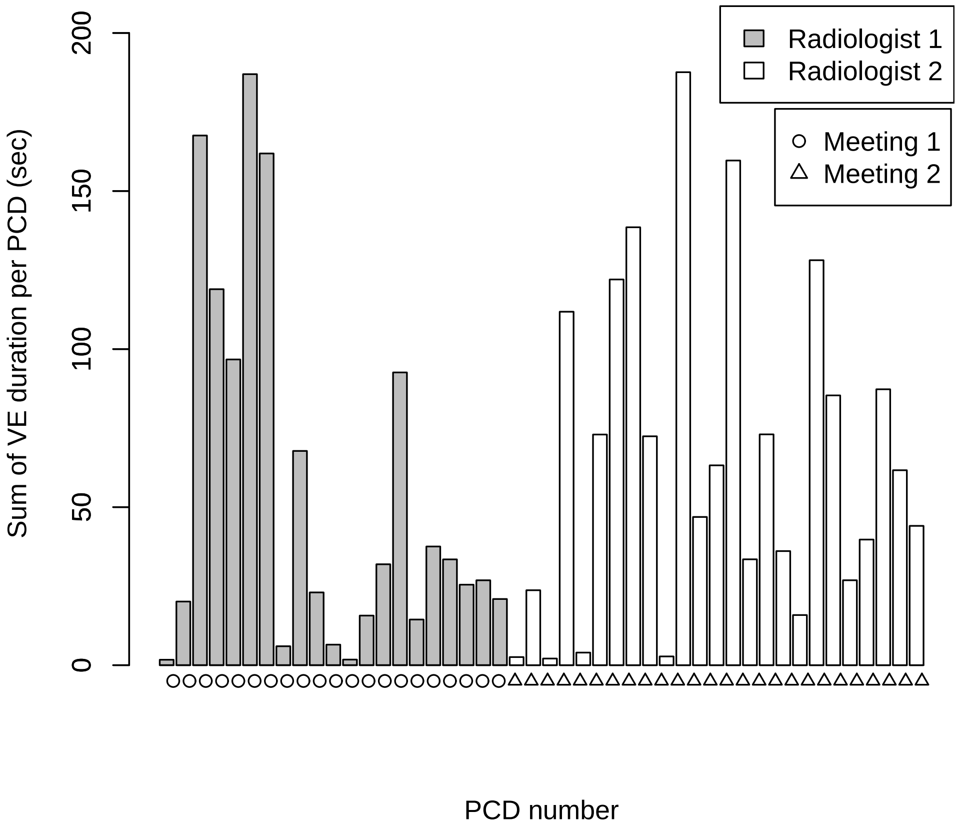


Figure 7.5: VE duration plot of meeting participants with the same role

speaker role instead of speaker IDs, in order to emphasise the function of a role but not of a person.

For example, Figure 7.5 shows the variation of VE duration per PCD by two radiologists in two meetings. In meeting 1, only Radiologist 1 (R1) participated, and he spoke in each PCD. His talk in each PCD has mean duration of 55.1 second with standard deviation 58.7 second. In meeting 2, R1 is absent and R2 took his role. The talk from R2 has mean duration of 65.7 second with standard deviation 51.7 second. So, on the perspective of VE duration per PCD, R1 and R2 are similar. I prefer a role variable "Radiologist" instead of their IDs, because any ID may be absent in a meeting and hence alter the regression or classification model. But a role variable is stable through each meeting without a serious modification on VE distribution. As a consequence, *Role* is a more comprehensive feature than $SpeakerID$ in MDTMs, and will be used in following experiments.

7.2 Fitting Statistical Models

In this study, our task is to predict if a vocalisation event is a topic start or not. A linear regression model can be constructed to specify the relations between $TopicStart$ and predictors which are vocalisational features automatically extracted from MDTM vocalisation events. Since $TopicStart$ is a binary variable, which violates the homogeneity of variance assumption of ordinary regression models, it cannot serve as a response in linear regression model. If I assume that the probability form $\pi(x)$ has a linear relation with x, the linear probability model (Equation 7.1) can be used to predict the probability with selected intercept α and slope β. But Equation 7.1 must satisfy that $\pi(x) > 0$ and $\pi(x) < 1$, so the value of x is bounded within a certain range and it makes the linear probability model difficult to use and to be interpreted.

$$\pi(x) = \alpha + \beta x \tag{7.1}$$

I apply logistic regression model [Agresti, 2002] to avoid the drawbacks of using a linear probability model. In this model, the value of x is not bounded.

Equation 7.2 shows that a log transformed fraction of success ($\pi(x)$) over failures (1-$\pi(x)$) has linear relation with x. The fraction part is called $Odds$. From this equation, a logistic regression model is not only used to predict success probability $\pi(x)$, but also capable of predicting the class of an instance. If $\pi(x) > 50\%$ then $logit[\pi(x)] = log(\frac{\pi(x)}{1-\pi(x)}) > 0$. So if the linear regression score from the right side of Equation 7.2 is positive, this instance is classified as positive class. Logistic regression is actually a classification model.

$$logit[\pi(x)] = log(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x \tag{7.2}$$

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{7.3}$$

Equation 7.3 is a transformation of Equation 7.2, for the convenience of predicting probability $\pi(x)$. The probability value is delivered by an exponential function of x. Logistic regression differs from ordinary linear regression in that it is a non-linear model over predictor and response, and consequently there is no need for predictors to be normally distributed, or have equal variance in each group. With these properties, a logistic model imposes less limitations on predictors.

$$p(Y_i = 1|X) = \frac{\exp(u)}{1 + \exp(u)}, \quad u = A + B_1X_1 + B_2X_2 + ... + B_nX_n \tag{7.4}$$

In order to predict the probability of a VE as $TopicStart$, with more than one vocalisational features, logistic regression model is adapted to accommodate more than one predictor (Equation 7.4), where $p(Y_i = 1|X)$ (noted as p_i) is the estimated posterior probability of the ith sample to be in class 1 and u is the ordinary multiple linear regression model. I apply Equation 7.4 on the MDTM corpus, to assess the possibility that an instance (VE) is a topic boundary. In order to determine the most useful features, I use the 'descending' selection method in model selection (Section 7.2.1).

7.2.1 Model Selection

Multiple Logistic Regression is selected as a model for binary response prediction. In this section I discuss the methods to determine the most important predictors and to determine the coefficients of predictors. In the end, a validation procedure is carried out on the fitted logistic model.

The 'descending' selection method in model selection has been employed. This means that all possible independent variables are included in the model at first, fitting the model against data, and then the least significant variable is removed. This procedure is iterated until all the variables in the model are significant. This leads to the final regression model.

Problems arise in calculation when all 733 vocalisations are used to fit the model. The negative of the Hessian matrix is not positive definite, and the convergence is questionable. The problem resides in 'redundant' data entries. In 733 vocalisations, only 22 entries match Case Start=1. It happens that for one speaker role, all his/her vocalisations match Case Start=0, so this speaker is not useful to classify vocalisations, and is redundant to calculation. Therefore I need to remove that speaker. The remaining dataset has 395 entries, and variables *Role* include *Surgeon*, *Physician*, and *Oncologist*.

7.2.1.1 Role horizon effects

Model selection (predictor selection) is achieved on this reduced dataset. First, *log* transformed VE duration $ln(d)$, speaker role and role horizons are included as predictors. I find that, r_{-1}, r_{-2}, r_1, r_2 are all insignificant [1]. So the *Role* which talks before or after one Case Start does not act as an influential factor to predict the probability of Case Start.

Variable	DF	Chi-Square	p > Chi-Square
ln(d)	1	10.33	0.0013
Role	2	6.1	0.0474

Table 7.2: Likelihood Ratio Statistics

[1] The null hypothesis is that a variable has no influence to the logit score. F-test [Kutner et al., 2005] shows that the null hypothesis is not challenged by the presented data set ($p > 0.05$), so the effect of these variables is insignificant.

$$\chi^2 = -2LogL_N - (-2LogL_F) = -2Log(\frac{likelihood_N}{likelihood_F}) \tag{7.5}$$

In Equation 7.5 L_N is the likelihood of null model, L_F is the likelihood of fitted model.

Ordinary linear regression models minimise least square error to estimate the coefficient of a predictor. But since the logistic curve is not linear, the coefficient of a predictor is obtained from maximum likelihood estimation (likelihood is a conditional probability value $P(y|x)$). With large samples, the difference of -2 times *log* likelihood follows the Chi-square distribution (Equation 7.5, Agresti [2002]). Table 7.2 shows the Chi-square statistics of each predictor, which tests against the null hypothesis that at least one of the predictors' regression coefficient is not equal to zero in the model. For example, *Role* corresponds to degree of freedom as 2, $\chi^2 = 6.1$ and $p = 0.0474$. This result means that comparing the null model (containing no predictor) against the regression model which contains only *Role* as predictor, the difference is significant (because $p < 0.05$). So, speaker role and *log* transformed VE duration are essential factors for the logistic model.

Variable	DF	Estimate	Std. Error	ChiSq	p > ChiSq
Intercept	1	-4.2144	0.5124	67.6391	< 0.0001
ln(d)	1	0.7078	0.2203	10.3269	0.0013
Role 1	1	-0.4218	0.5136	0.6743	0.4116
Role 2	1	0.7788	0.338	5.3091	0.0212

Table 7.3: Multiple Logistic Regression model 1

$$logit(p_i) = -4.21 + 0.71 * ln(d) - 0.42 * Oc + 0.78 * Ph \tag{7.6}$$

$$p_i = \frac{\exp(-4.21 + 0.71 * ln(d) - 0.42 * Oc + 0.78 * Ph)}{1 + \exp(-4.21 + 0.71 * ln(d) - 0.42 * Oc + 0.78 * Ph)} \tag{7.7}$$

In Equation 7.6 and 7.7 d =duration, Oc =Oncologist, Ph =Physician

In logistic model (Table 7.3), *Role* is a categorical variable, it is beneficial to discover the effect of each level in Role. Role 5 (Surgeon) is treated as a reference category, so it is not presented as a covariant. For Role 2 (Physician), its effect is significantly different from Role 5 ($p < 0.05$), while Role 1 (Oncologist) is not. Other roles do not fit this model, because they refer to Case Start = 0 in the training set. The fitted logistic regression model is shown in Equation 7.7. In this equation, the intercept means that when an VE duration is 1 second, and the speaker role is Surgeon (the reference role), $logit(p_i) = -4.21$ and the $Odds = exp(-4.21) = 0.015$. In this case, $p(x = 1) < 1.5\%$. The intercept stands for the probability when the numerical predictor is zero and the categorical predictor is at reference level. The coefficient of $ln(t)$ is 0.71, which means a unit increase of $ln(t)$ corresponds to an $exp(0.71)$ increase on the $Odds$. Moreover, the coefficients of binary predictors Oc and Ph indicate the effect of speaker role on $Odds$ ratio. For example, if the speaker is an Oncologist instead of the reference Surgeon, the $Odds$ decreases by $exp(0.71)$. Generally, a PCD is most likely to start with a long VE and the speaker is Physician.

Variable	DF	Chi-Square	p > Chi-Square
r_{-1}	4	4.31	0.3657
r_{-2}	4	38.85	<0.0001
r_1	4	4.31	0.3657
r_2	4	38.85	<0.0001

Table 7.4: Chi-square statistics to test the correlation between *Role* and its horizon variables

In the 'descending' selection procedures, each of the *Role* horizon features is dropped from the logistic model because its effect is not significant. Since *Role* horizon is adapted from *Role*, I doubt if the reason is multicollinearity[1] between *Role* and horizon features. The Chi-square test is employed to test correlation between categorical variables. If $p < 0.05$, the independence assumption between two variables is rejected, hence they are correlated. Table 7.4 shows the signifi-

[1] Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.

cance of correlation. Only r_{-2} and r_2 significantly relate to *Role*. r_{-1} and r_1 are independent from *Role*. So I can not assert that categorical horizon variables always suffer from multicollinearity. The Chi-square test should be used to validate variable independency for each model.

7.2.1.2 VE duration horizon effect

I perform model selection method on *Role* horizons in Section 7.2.1.1. In this section I follow a descending selection procedure to test VE duration horizon as a novel feature for the logistic model. Horizon level 2 is applied to *log* VE duration, noted as $ln(d_{-1})$, $ln(d_{-2})$, $ln(d_1)$ and $ln(d_2)$. Speaker role and *log* transformed current VE duration are also included as candidates.

Variable	DF	Estimate	Std. Error	ChiSq	Pr > ChiSq
Intercept	1	-4.6854	0.6237	56.4275	< 0.0001
$ln(d)$	1	0.6449	0.2167	8.8551	0.0029
$ln(d_1)$	1	0.4989	0.2219	5.0555	0.0245

Table 7.5: Multiple Logistic Regression model 2

$$logit(p_i) = -4.69 + 0.64 * ln(d) + 0.5 * ln(d_1) \tag{7.8}$$

$$p_i = \frac{\exp(-4.69 + 0.64 * ln(d) + 0.5 * ln(d_1))}{1 + \exp(-4.69 + 0.64 * ln(d) + 0.5 * ln(d_1))} \tag{7.9}$$

The MLR model 2 with VE horizon features are presented in Table 7.5 and Equation 7.8 and 7.9. This model contains two numerical predictors: *log* VE duration and *log* duration of the next VE. Both predictors are significant in this model.

Similarly to MLR model 1 (Table 7.3), the intercept in model 2 stands for $Odds = exp(-4.69)$ when $d = 1$ and $d_1 = 1$. $ln(d)$ and $ln(d_1)$ each have positive effect on the base *Odds* value. If the current VE is relatively long, and the closely preceding VE is also long, a PCD has high probability to start on the current

VE.

Variable	N	Correlation Estimate	p
$ln(d_{-1})$	394	0.114	0.023
$ln(d_{-2})$	393	0.106	0.036
$ln(d_1)$	394	0.114	0.023
$ln(d_2)$	393	0.106	0.036

Table 7.6: Pearson correlation test between VE duration $ln(d)$ and its horizon variables

In addition to MLR model selection, I am interested in the correlations between VE duration horizon features. Pearson correlation statistics is used to calculate correlation between numerical variables. Correlation ρ_{xy} is the covariance between x and y divided by the root square of variance product of x and y (Equation 7.10). When x and y are highly correlated, ρ_{xy} approaches 1.

$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{E[(x-E(x))(y-E(y))]}{\sqrt{E[x-E(x)]^2 E[y-E(y)]^2}} \tag{7.10}$$

Correlation deserves attention in this case because where x and y are jointly normally distributed, $\rho_{xy} = 0$ implies that x and y are independent ([Neter et al., 1996], pp.641). Since $\rho_{xy} = 0$ is very rare, I am more interested in a general criterion to judge independence from Pearson correlation. Neter indicated that if the independence assumption holds, t^* follows the t distribution with (n-2) degrees of freedom (Equation 7.11, where n is the number of samples, r_{xy} is sample correlation).

$$t^* = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \tag{7.11}$$

Table 7.6 shows $p < 0.05$ for each entry, so correlation does not follow the t distribution and the independence assumption for VE duration $ln(d)$ and its horizon variables are violated. I need to consider the interaction between $ln(d)$ and $ln(d_1)$ in MLR model 2. The interaction is presented as a product of $ln(d)$

and $ln(d_1)$. A modified version (model 3) is shown in Table 7.7 and Equation 7.12, where each covariant is significant. In model 3, the coefficients of $ln(d)$ and $ln(d_1)$ are higher than in model 2, but the interaction term $ln(d) * ln(d_1)$ has negative effect on $Odds$. In Section 7.2.2 I compare and validate each model.

Variable	DF	Estimate	Std. Error	ChiSq	Pr > ChiSq
Intercept	1	-6.4444	1.0402	38.3786	< 0.0001
$ln(d)$	1	1.5496	0.4148	13.9528	0.0002
$ln(d_1)$	1	1.4497	0.4248	11.6486	0.0006
$ln(d) * ln(d_1)$	1	-0.5238	0.2007	6.8121	0.0091

Table 7.7: Multiple Logistic Regression model 3

$$logit(p_i) = -6.44 + 1.55 * ln(d) + 1.45 * ln(d_1) - 0.52 * ln(d) * ln(d_1) \quad (7.12)$$

7.2.2 Model Validation

In the preceding section, I fit a logistic regression model to predict the probability of a vocalisation event as the start of a PCD from its corresponding vocalisation features. Chi-square statistics is applied to test the importance of each factor, but I need another metric to validate the goodness of fit of the predicted model toward the data set.

$$\chi^2_{HL} = \Sigma^g_{i=1} \frac{(\Sigma_j y_{ij} - \Sigma_j \widehat{\pi}_{ij})^2}{(\Sigma_j \widehat{\pi}_{ij})[1 - (\Sigma_j \widehat{\pi}_{ij})/n_i]} \quad (7.13)$$

The Hosmer-Lemeshow (HL) test [Hosmer and Lemeshow, 1980] is a modified Pearson statistic. HL ranks fitted samples by probability and groups the samples into equal size partitions, then compares the observed and fitted counts for each partition. Equation 7.13 is the HL statistic, in which y_{ij} denotes the binary score of observation j in group i, and $\widehat{\pi}_{ij}$ denotes the corresponding predicted probability of the logistic model over ungrouped data. When the number of distinct patterns of covariate values (for the original data) is close to the sample size, the null distribution is approximated by Chi-squared with df = number of groups -2 ([Hosmer and Lemeshow, 2000], pp.147-156). Table 7.8 presents the partition of samples as 10 groups and the expected sample frequency in each group. The *Observed* and *Expected* values from model 1 are quite close to each other, hence its Chi-square score 5.74 is relatively low with $p > 0.05$. HL test concludes that model 1 fits data well.

		caseB	=1	caseB	=0
Group	Total	Observed	Expected	Observed	Expected
1	40	0	0.45	40	39.55
2	40	0	0.63	40	39.37
3	40	2	0.86	38	39.14
...	...	...	...	...	...
8	40	2	2.87	38	37.13
9	40	4	4.21	36	35.79
10	35	7	6.57	28	28.43

Table 7.8: Partition for the Hosmer and Lemeshow test in Model 1

	Chi-Square	DF	p > ChiSq
model 1	5.7447	8	0.6758
model 2	6.4961	8	0.5918
model 3	18.6297	8	0.017

Table 7.9: Hosmer and Lemeshow Goodness-of-Fit test on each MLR model

Table 7.9 shows the result of Goodness-of-fit test for each MLR model, where HL statistics score is evaluated against Chi-square distribution. p values show that the null hypothesis of Chi-square distribution is not violated by model 1 or model 2, but model 3 suffers from lack-of-fit. At this stage, I can conclude that model 1 and 2 are successful logistic regression model to predict topic boundary probability. But problems emerge in their classification consequences. As a logistic regression model predicts a probability value (Equation 7.3) from a sample's features, it simultaneously aligns this sample into a positive or negative class (in Equation 7.2, $\pi(x) > 0.5$ then $logit[\pi(x)] > 0$, the sample is classified as positive). However, the classification results (Table 7.10) are disappointing, as no one positive sample is predicted. This result challenges the validity of the Hosmer and Lemeshow test.

Table 7.10: Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c)

(a)

a	b	← pred
0	22	a=1
0	373	b=0

(b)

a	b	← pred
0	22	a=1
0	373	b=0

(c)

a	b	← pred
0	22	a=1
0	373	b=0

Revisiting HL test result is in Table 7.8, I see that all the samples are sorted in descending order by their predicted probability of being class 0. In each group, the *Expected* value is the sum of all predicted probability values in this group (Equation 7.13). The expected value is quite close to observations in each group, then it results in a relatively low χ^2_{HL} score at $DF = 8$. As a consequence, the difference between expectation and observation is not significant.

The question is: since the expected probabilities are quite close to observations, why does MLR classify all samples as negative? The answer could be on two sides. First, MLR predicts probability $p(Y_i|X)$ with high accuracy. The linear regression part of Equation 7.2 is effective. Second, $p(Y_i = 1|X) < 0.5$ is recorded for each sample as class 1. Since 50% is the default threshold, all samples are classified as negative. There are only 22 positive out of 395 samples in the dataset, the proportion of positive samples is only 5.6%. Since MLR coefficient β is obtained through maximum likelihood estimation, the predicted probability value $p(Y_i = 1|X)$ is expected to be low.

In order to increase class prediction accuracy, I can adjust the probability threshold for the positive class, instead of altering the probability values from MLR model. An arbitrary threshold 15% is proposed and consequently the $Odds = \frac{0.15}{1-0.15} = 17.6\%$, and $logit = \log(0.176) = -1.73$. Therefore when the linear regression part of any MLR model predicts $logit > -1.73$, this sample is classified as positive. Three predicted MLR models are adapted with this new probability threshold and keep all other parameters fixed. I show a new confusion matrix for each of them in Table 7.11. Each model has improved classification accuracy on class 1. All MLR classification experiments above are performed on training set, in order to validate the models, I test them with a new dataset, which contains 535 VEs and 25 topic boundaries. The speakers and roles in this dataset are presented in Table 7.17. Since no *Physician* participated in this meeting, *Ph* factor will be zero for each instance. I see that with a 15% threshold the model becomes prone to predict false positives (Table 7.12). Since this threshold is arbitrary, I apply an adaptive threshold in the test set and aim to balance FP and FN. Classification confusion matrix is shown in Table 7.13.

I will discuss segmentation accuracy of MLR models in the following section. Model 3 was tested to be lack of fit, but since each factor in it has significant effect on predicted probability, I retain model 3 and evaluate it with segmentation metrics.

Table 7.11: Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c), with positive threshold 15% (in training set)

(a)

a	b	← pred
4	18	a=1
19	354	b=0

(b)

a	b	← pred
2	20	a=1
17	355	b=0

(c)

a	b	← pred
5	17	a=1
24	348	b=0

7.2.3 Segmentation performance

In this section, I focus on segmentation performance of three MLR models. MLR sets classification threshold to be 50% by default, which results in no topic boundary prediction in training set. This is induced by highly unbalanced dataset in two classes. I apply 15% threshold instead and have good segmentation accuracy

Table 7.12: Confusion Matrix of predictions from logistic regression model 1 (a), model 2 (b) and model 3 (c), with positive threshold 15% (in test set)

(a)

a	b	← pred
1	24	a=1
7	503	b=0

(b)

a	b	← pred
7	18	a=1
63	447	b=0

(c)

a	b	← pred
6	19	a=1
45	465	b=0

Table 7.13: Confusion Matrix of predictions from logistic regression model 1 (threshould=10%) (a), model 2 (thresold=30%) (b) and model 3 (thresold=20%) (c) (in test set)

(a)

a	b	← pred
1	24	a=1
17	493	b=0

(b)

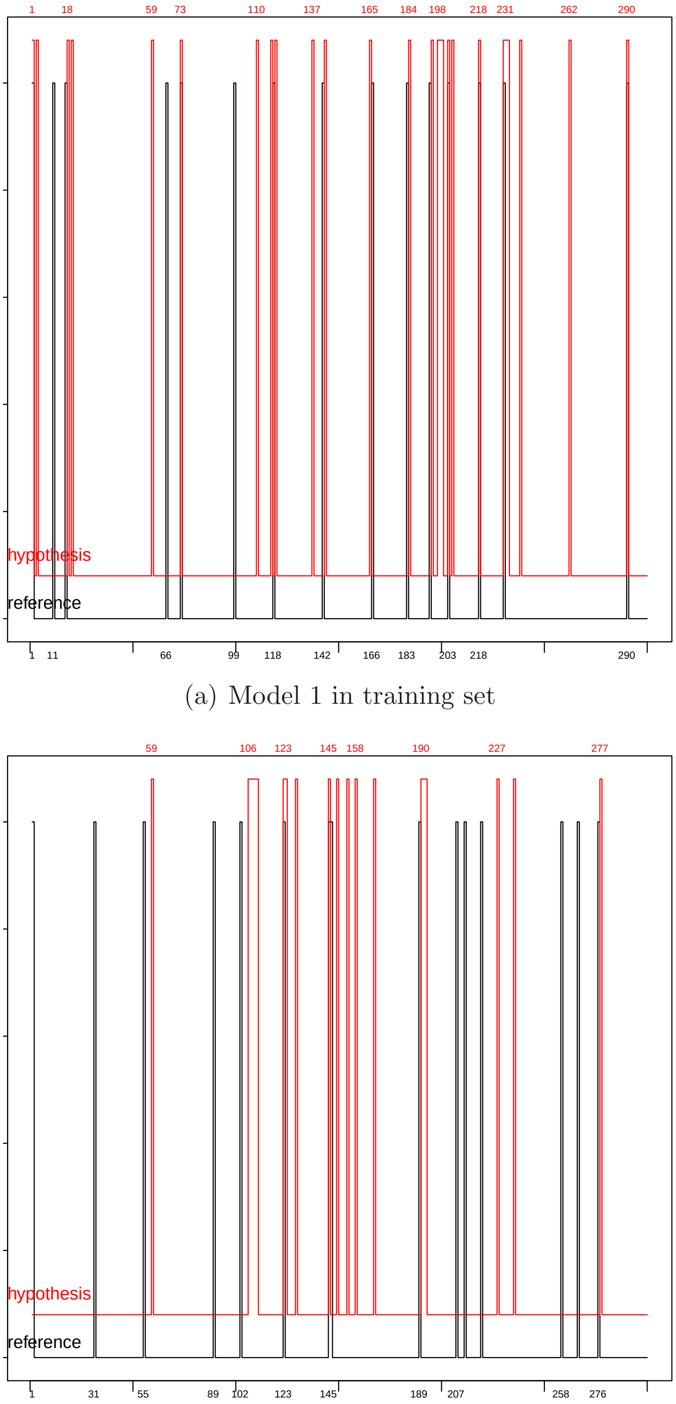
a	b	← pred
4	21	a=1
21	489	b=0

(c)

a	b	← pred
3	22	a=1
27	483	b=0

(Table 7.14). But 15% arbitrary threshold is not proper for test set, because it produces either too few or too many boundaries which deviates ω from 1 (Table 7.15). I then apply 10%, 30% and 20% thresholds for each model, for $\omega \to 1$ (Table 7.16). Under this setting, P_k and WD better indicate the effect of classifiers.

Model 1 performs best with the training set (in Figure 7.6(a), the predicted boundaries mostly match the reference), but it is weak for a test set missing certain roles. Model 2 shows better adaptivity, since it only has VE duration and duration horizon covariants (Figure 7.6(b)). Multiple logistic regression model confirms the effect of VE horizon on duration instead of roles. Model 3 is not



(a) Model 1 in training set

(b) Model 2 in test set

Figure 7.6: Segment boundary plot of multiple logistic regression models

Table 7.14: Segmentation accuracy of three Logistic Regression models (in training set)

	P_k	WD	ω
model 1	**0.276**	**0.393**	**1.045**
model 2	0.358	0.487	0.864
model 3	0.316	0.477	1.318

Table 7.15: Segmentation accuracy of three Logistic Regression models (in test set with 15% threshold)

	P_k	WD	ω
model 1	0.444	0.505	0.32
model 2	0.448	0.541	2.8
model 3	0.438	0.564	2.04

successful either with Hosmer and Lemeshow test or segmentation metrics, so the interaction term $ln(d) * ln(d_1)$ between VE duration and its horizon is not recommended. Generally, I should mention the drawbacks of MLR on topic segmentation. MLR has reliable probability prediction for instances, but it requires an adaptive threshold for classification, which determines segmentation accuracy. I have not found a robust method to produce the threshold and presently I need arbitrary trials for a classification result satisfying $\omega = 1$.

7.2.4 Conclusion

In order to discover the relation between vocalisational features and the probability of a VE as PCD leading talk, I analyse variants of regression model, and find that multiple logistic regression model (MLR) is adequate to present a linear relation between (logit format) probability and features. Moreover, MLR has less stringent requirements on predictor's distribution.

Various vocalisation features are involved in the model selection procedure. *log* transformed VE duration is preferable because its distribution is closer to normal. Speaker role is selected over speaker ID because the former better represents MDTMs routine and structure. The descending model selection method removes most horizon features and retains only $ln(d_1)$. The MLR model only

Table 7.16: Segmentation accuracy of three Logistic Regression models (in test set)

	Threshold	P_k	WD	ω
model 1	10%	0.507	0.558	0.72
model 2	30%	**0.364**	**0.406**	**1.0**
model 3	20%	0.419	0.497	1.2

includes the features with significant effect on *log* Odds, and a successful MLR model must fit the data set well.

Finally two MLR models are presented in Equation 7.7 and 7.9. A MLR model is derived from maximum likelihood estimation, the *logit* form probability of PCD start VE is precisely estimated from VE duration and speaker role, or VE duration and its horizon. Furthermore, the Hosmer-Lemeshow test validates that the predicted MLR model fits the distribution of dataset and can be used for topic boundary prediction.

On the other hand, the unbalanced dataset seriously harms MLR classification accuracy with a default 50% threshold. I adapt this threshold to an arbitrary lower level and get satisfying segmentation accuracy with Model 2 on test set. As a conclusion, multiple logistic regression model is a competitive approach to predict topic boundary with content-free features, but it is still an open question how to determine a robust threshold for positive class.

Topic boundary detection is naturally a classification problem, with the aim to separate topic boundary vocalisations from non-boundary ones. In order to avoid the influence of arbitrary probability threshold on classification output, I study more classification approaches in Section 7.3, including naïve Bayes classifier and ensemble classifier. I aim to find a robust classification model on unbalanced dataset and achieve automatic topic segmentation.

7.3 PCD segmentation through classification

With the AMI corpus, I validated the value of vocalisation Horizon as data representation for segmentation. I also tested variations of naïve Bayes classifiers, CRF and ensemble classifiers. From those experiments, I formed an assessment of which types of classifiers fit unbalanced and sequentially correlated data. The most important achievement in the AMI corpus is that I built a framework for evaluating the goodness of topic segmentation, which included P_k, WD and balance factor ω. At this stage, I am curious about how reliably this framework can be applied to the MDTM data set and if there is difference between simulated meetings (AMI) and real ones (MDTM). In this section I will study MDTM segmentation toward a conclusion.

7.3.1 MDTMs vocalisation unit and features

I sample MDTMs speaker talk with similar methods as in the AMI corpus. Instead of using a fixed length sliding window, I generate MDTMs classification instances based on complete vocalisation events (VE) and extract features for each VE. For the AMI corpus, I discussed two types of vocalisation sampling methods in case of overlapping (Section 3.1.2), where VE_t terminates at the beginning of an overlap and VE_c terminates at the natural ending of one vocalisation from a speaker. In MDTMs, VE_t is used as the standard of vocalisation event annotation, because it is difficult to locate the accurate end time of a VE during intense group talk. Manually annotated vocalisation events act as reference for speaker segmentation and topic boundary instances classification.

Another aspect worth noticing is filled pause (Section 3.1.2.2). I have successfully extracted filled pauses in the AMI corpus (Section 3.1.2) and tested its utility with various classification schemes (Section 6.3.1 and 6.2.4). Since FP_s (Section 3.1.2) does not show clear advantage over EP_s, and MDTMs are not annotated on word-level, I no longer test filled pause effect in MDTMs. VE_t is extracted only with empty pauses. In this section, I test various features and determine the most favorable combination of classification feature sets in the end.

Comparing with the AMI corpus of 30 meetings, the annotated MDTM corpus only contains 2 meetings, each is shorter than 2 hours. So in each classification model I use 5-fold cross validation. Just for the AMI corpus, the n-fold training and testing set are not partitioned randomly, but sequentially. The instances in test set must follow time order. This setting is useful to keep relatively equal quantity of topic boundaries allocated in each fold of test set. The PT NB

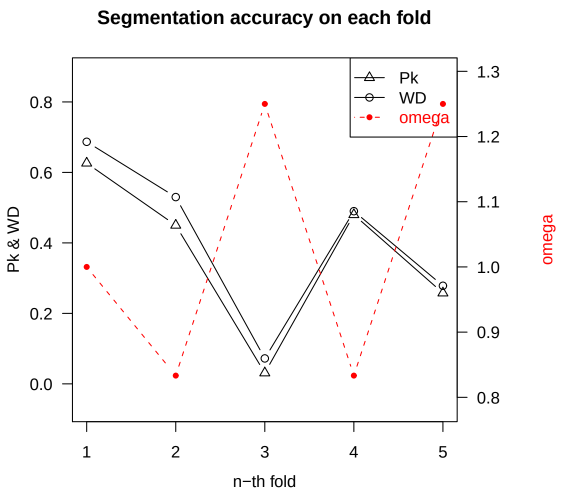


Figure 7.7: Segmentation accuracy on each fold (5-fold PT NB classification with EP_c features)

segmentation accuracy with EP_c features (Section 7.3.1.3) is presented in Figure 7.7. Although P_k and WD varies among folds, ω is relative stable between 0.8 and 1.3. The number of reference boundaries in each fold is (5,6,4,6,4), and the number of predicted boundaries is (5,5,5,5,5). So our dataset partition setting is reliable. In next subsections I will discuss feature settings.

7.3.1.1 Using speaker ID or speaker role

In the MDTMs corpus, it is common that more than one medical persons of the same role take part in one meeting. Each of them is in charge of certain patients, but they participate in the discussion of each patient. The role definition of each MDTM participant is in Table 7.17.

The basic assumption is that people who share common roles will behave similarly in medical team meetings [Kane and Luz, 2009]. A role represented format may bring a more general feature for topic boundary detection.

In order to test this hypothesis, *Speaker ID* and *Role* are used separately on topic boundary classification. PT-NB is one of the best classifiers on the AMI

Roles	Speaker ID
Pathologist	p1
Surgeon	s1, s2
Oncologist	o1
Radiation Oncologists	ro1
Clinical Oncologists	co1
Radiologist	r1, r2
Medical Consultants	mc1, mc2, mc3, mc4, mc5, mc6, mc7
Medical Registrars	mr1, mr2, mr3
Surgical Registrars	sr1
Physician	Null
Nurse	n1
Junior Members	jm1, jm2

Table 7.17: Speaker Roles and corresponding speaker IDs (for the consideration of privacy, speaker ID is replaced by the role code)

corpus, I use it again on the MDTMs corpus. Topic segmentation accuracy with 5-fold cross-validation and adjacent prediction filtering (NoAdj) is shown in Table 7.18. We can see that, *Role* performs better than *Speaker ID* with each metric. The possible reason is that speaker ID has higher variability than roles, and this level of variability tunes NB classifier positively. In MDTMs topic segmentation experiments, I use roles instead of speaker ID.

Table 7.18: Segmentation accuracy of PT-NB with discrete single feature of MDTMs data

	P_k	WD	ω
Speaker ID	0.354	0.538	2.24
Role	0.273	0.481	2.0

7.3.1.2 Transforming vocalisation duration

The distribution of vocalisation duration from MDTMs is highly right-skewed (Figure 7.2), it has $skewness = 4.45$ in this dataset. Serious non-symmetry undermines normality of continuous variables. In order to minimize the influence of skewness, Log transform is performed on vocalisation duration. The distribution of Log(VE_duration) has $skewness = 0.79$ (Figure 7.4).

The effect of Log transform with topic boundary classification is shown in Table 7.19. Contrary to our expectation, Log transform leads to worse ω and P_k than the original feature. The possible reason is that transformed duration results in less variability, and decreases the likelihood estimation on topic boundary instances. So I select VE duration as one feature for classification experiments instead of its Log transformed values.

Table 7.19: Segmentation accuracy of PT-NB with continuous single feature of MDTMs data

	P_k	WD	ω
VE_duration	0.427	0.509	1.0
Log(VE_duration)	0.43	0.478	0.92

7.3.1.3 feature sets

In the AMI corpus, Vocalisation Horizon achieves better segmentation accuracy than original features with certain classifiers. I extend horizon concept to MDTMs and modify the horizon features. Since *Pause* does not show significant effect on segmentation accuracy, I drop it from feature sets. Here *SpeakerRole* is included as another horizon feature (as Equation 7.16 and 7.17).

In Equation 7.14, EP_c includes most basic VE information: VE start time (st), VE duration (d) and speaker role (r). In Equation 7.15, EP_D includes all features in EP_c plus VE duration horizon d_n, which is the VE duration of n vocalisations preceding and following current VE. Equation 7.16 stands for feature set EP_R which contains basic features EP_c plus speaker role horizon. The last feature set EP_{DR} is a combination of EP_D and EP_R. Different horizon features may lead to various classification accuracy. The combination feature set is also an indicator of feature complexity in classification.

$$\begin{aligned}
EP_c &= (r, st, d) && (7.14)\\
EP_D &= (r, st, d, d_{-n}, ..., d_{-1}, d_1, ..., d_n) && (7.15)\\
EP_R &= (r, st, d, r_{-n}, ..., r_{-1}, r_1, ..., r_n) && (7.16)
\end{aligned}$$

$$EP_{DR} = (r, st, d, d_{-n}, ..., d_{-1}, d_1, ..., d_n, r_{-n}, ..., r_{-1}, r_1, ..., r_n) \quad (7.17)$$

Table 7.20: Notations of Equation 7.14 to 7.17

Notation	Description
r	speaker role
st	start time of vocalisation event (VE)
d	VE duration
d_{-n} to d_n	Horizon level n of VE duration
r_{-n} to r_n	Horizon level n of speaker role

Luz [2009] stated that horizon level 3 and 5 performed best on P_k and WD separately, so I take horizon level $n = 3$ for simplicity. Moreover, I include an extra feature *st*, the start time of VE, in each feature set. Since I assume vocalisation horizon indicates 'local' information of adjacent VE, *st* may indicate the 'global' sequence of VE. Two adjacent VEs have low probability to be boundaries both. Table 7.21 shows that EP_c has minor advantage over the one without *st*, so I keep *st* as one potential factor.

Table 7.21: PT-NB segmentation accuracy of EP_c and EP_c without *st*

Feature Set	P_k	WD	ω
EP_c - st	0.4	0.465	1.08
EP_c	0.394	0.442	1.0

7.4 Segmentation experiments

In this section, successful classifiers from the AMI corpus are tested on MDTMs data, with simple feature set and horizon based feature sets. The classifiers include thresholded naive Bayes classifiers (Section 7.4.1) and Ensemble classifiers (Section 7.4.2). I use P_k, WD and ω to evaluate segmentation accuracy and analyse the effect of features together with classifiers.

Table 7.22: Segmentation accuracy of PT-NB with MDTMs

	P_k	WD	ω
EP_c	0.394	0.442	**1.0**
EP_D	0.402	0.434	0.76
EP_R	**0.32**	**0.402**	0.92
EP_{DR}	0.377	0.417	0.72

7.4.1 naïve Bayes classifier

PT-NB is tested to be an effective classifier for AMI topic segmentation. Consequently, it is used as the primary classifier for MDTMs data set. 5-fold cross-validation results of PT-NB with different Horizon settings are shown in Table 7.22. Role horizon feature set EP_R presents the highest P_k and WD scores with the second best ω. Other horizon types are worse than EP_R in each metric. Simple features EPc has nearly the worst P_k and WD scores although its $\omega = 1$. This experiment validates the role horizon effect on MDTMs PCD segmentation against simple features as well as other horizon types. Table 7.23 shows that EP_c generates no positive predictions on the full test set with MAP NB classifier. In this case, I are convinced that simple feature set, such as EP_c, is too weak to produce prominent likelihood ratio for features of positive instances in training set, and hence the posterior distribution of simple features is unlikely to be overwhelming for certain instances. As a consequence, EP_c has weak prediction power with MAP NB.

Revisiting Table 7.23, I notice that EP_{DR} is superior over EP_c, especially with ω, which illustrates the importance of VE duration and role horizon for segmentation. Horizon features are essential for naïve Bayes classifiers. On the other hand, the difference of EP_c segmentation accuracy between MAP NB and PT NB signifies the advantage of naïve Bayes classifier and proportional threshold. Although the values of predicted probability are lower than 50%, the reference instances (TP) generally has higher probability value than others.

Table 7.23: Comparison of horizon effect with various naïve Bayes classifiers

classifier	features	P_k	WD	ω
MAP NB	EP_c	0.417	0.455	0
FT NB $\star$	EP_c	0.417	0.455	0
PT NB	EP_c	0.394	0.442	1.0
MAP NB	EP_{DR}	0.274	0.36	1.28

$\star$ threshold probability p=0.99

7.4.2 Ensemble classifier

Ensemble classifiers are designed to learn more expressive concepts than a single classifier, and they exhibit good segmentation accuracy with C4.5 base classifier in the AMI corpus. MDTMs are annotated from real meetings and there are more participants than AMI environment, I test ensemble classifier with C4.5 and MAP naïve Bayes classifier separately.

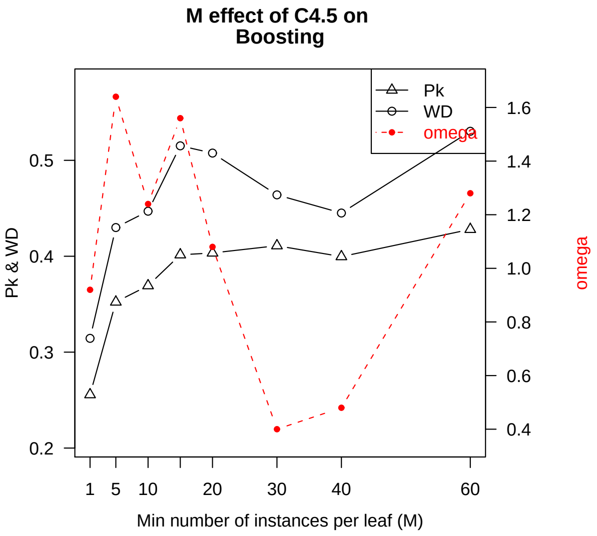
7.4.2.1 C4.5 base classifier

In Bagging and Boosting algorithms, the number M (Min number of instances per leaf) controls the complexity of C4.5 base classifier and influences ensemble classifier accuracy. Figure 7.8 illustrates the relations between M and P_k, WD, ω in Bagging and Boosting, where M is set from 1 to 60. From both figures, $M = 1$ is almost the best setting for P_k, WD and ω. I will use $M = 1$ for horizon features.

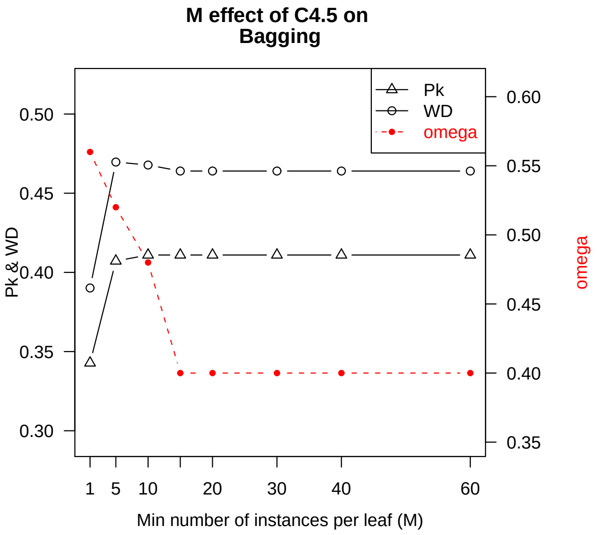
Table 7.24: Segmentation accuracy of AdaBoostM1 with C4.5 base classifier ($M = 1$)

	P_k	WD	ω
EP_c	**0.256**	**0.314**	**0.92**
EP_D	0.337	0.413	0.92
EP_R	0.314	0.375	0.88
EP_{DR}	0.368	0.43	0.76

In Table 7.24, EP_c reaches the best segmentation accuracy so far. Simply within the AdaBoostM1 (on C4.5 base classifier) experiments, EP_c performs better than other horizon feature sets. Boosting favors simple feature sets of MDTM



(a) AdaBoostM1



(b) Bagging

Figure 7.8: Accuracy of ensemble classifiers on EP_c features with various C4.5 leaf settings

Table 7.25: Segmentation accuracy of Bagging with C4.5 base classifier ($M = 1$)

	P_k	WD	ω
EP_c	0.343	**0.39**	0.56
EP_D	0.362	0.434	**0.84**
EP_R	**0.337**	0.398	0.56
EP_{DR}	0.423	0.474	0.48

corpus. A portion of EP_c boundary prediction plot is shown in Figure 7.9(a). In this figure, most predicted boundaries are quite close to real positions, although there are several missed ones.

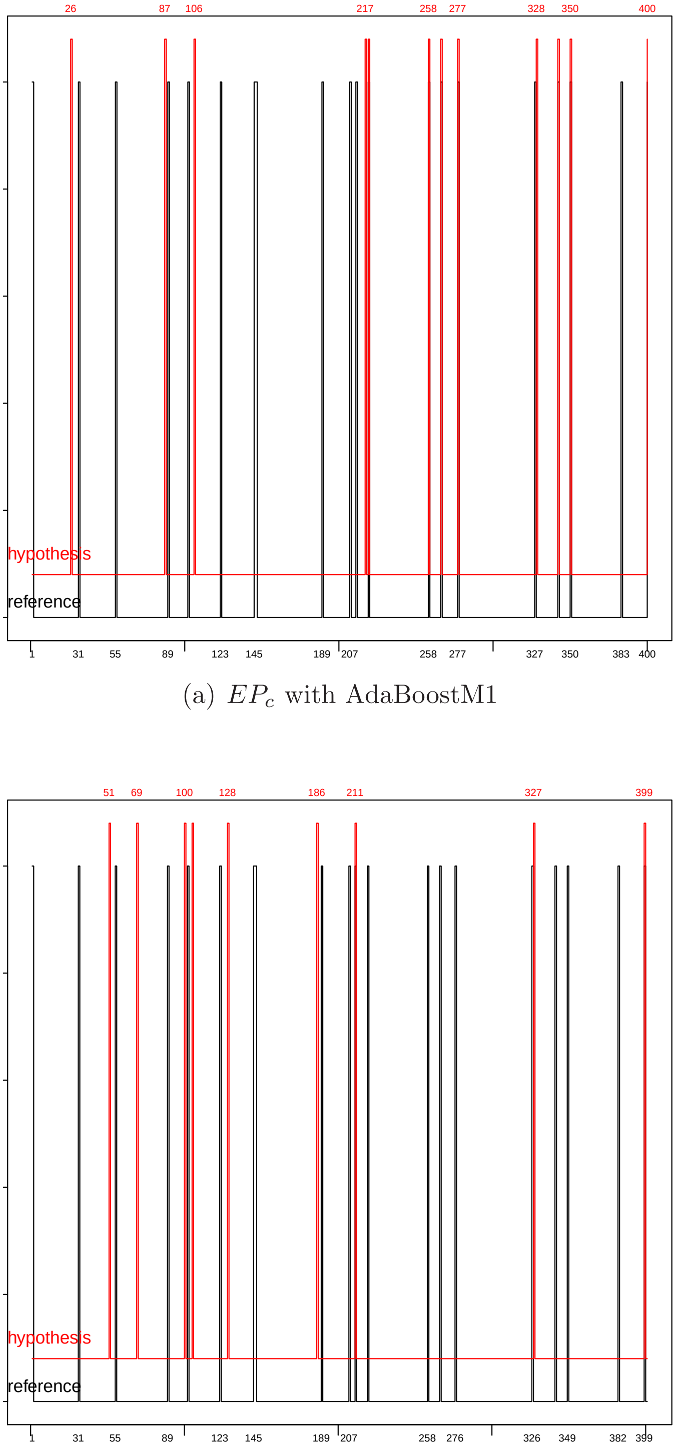
On the contrary to Boosting segmentation results, Bagging does not present an optimal choice from various features sets (Table 7.25). EP_R and EP_c has best P_k and WD separately, but EP_D has the best ω. I prefer EP_D for the quantity of its boundaries is most close to reference. Figure 7.9(b) shows more missing predictions than Figure 7.9(a), but its predicted boundaries are mostly accurate.

Table 7.26: Segmentation accuracy of C4.5 unpruned decision tree ($M = 1$)

	P_k	WD	ω
EP_c	0.318	0.486	2.0
EP_D	**0.251**	**0.335**	**1.04**
EP_R	0.349	0.434	1.24
EP_{DR}	0.413	0.493	1.04

In order to evaluate the effectiveness of Bagging and Boosting over its base classifier, I test C4.5 unpruned decision tree independently for MDTMs (Table 7.26). C4.5 classifier has competitive performance over Bagging and Boosting. On EP_D features, C4.5 has better P_k and WD score than Bagging and Boosting. Generally C4.5 has a higher ω score, and horizon feature sets have ω close to 1.

I propose Boosting with EP_c and C4.5 classifier with EP_D as optimal choices in this section. Experiments show that the basic decision tree classifier performs well with VE duration horizons. Nevertheless, Boosting has limited improvements on base classifier and it only significantly improves the prediction of EP_c. Bagging reduces ω score on each feature set, but it results in a too low ω. As a conclusion, C4.5 shows advantage with MDTM horizon features, and the Boosting algorithm



(a) EP_c with AdaBoostM1

(b) EP_D with Bagging

Figure 7.9: Plot of predicted boundaries from winning feature sets in Boosting and Bagging(C4.5 base classifier)

has positive impact only with simple feature set.

7.4.2.2 MAP naïve Bayes base classifier

In this section, I use MAP naïve Bayes base classifier for Bagging and Boosting, instead of C4.5 decision tree. Since naïve Bayes classifier performs well as independent classifier on the topic segmentation task, I expect the ensemble methods will improve segmentation accuracy, and avoid the measures to control base classifier complexity.

Table 7.27: Segmentation accuracy of MAP naïve Bayes classifier

	P_k	WD	ω
EP_c	0.417	0.455	0
EP_D	**0.272**	0.37	1.6
EP_R	0.398	0.465	**0.96**
EP_{DR}	0.274	**0.36**	1.28

Table 7.28: Segmentation accuracy of AdaBoostM1 with MAP naïve Bayes base classifier

	P_k	WD	ω
EP_c	0.415	0.466	0.44
EP_D	**0.286**	**0.368**	**1.2**
EP_R	0.434	0.512	1.4
EP_{DR}	0.297	0.368	0.64

Table 7.29: Segmentation accuracy of Bagging with MAP naïve Bayes base classifier

	P_k	WD	ω
EP_c	0.415	0.453	0
EP_D	0.282	0.436	1.4
EP_R	0.394	0.465	0.48
EP_{DR}	**0.276**	**0.379**	**1.04**

In order to evaluate ensemble classifier performance, I compare Bagging and Boosting separately with an independent MAP naïve Bayes classifier. Table 7.27

shows NB segmentation accuracy with 4 feature sets. In this table, EP_c corresponds to very low ω scores, and should not be used as useful guidance to meeting audience. EP_{DR} presents moderate accuracy on each metric and becomes the best choice for MAP NB. These facts validate the effect of horizon on certain features and disclose the deficiency of very simple feature set EP_c, which has low prediction power for unbalanced MDTMs data.

Boosting performs weighted learning over the whole training set, and increases the weight on difficult cases. Comparing AdaBoostM1 (Table 7.28) and MAP NB (Table 7.27) with P_k and WD, I see that EP_c, EP_R and EP_{DR} do not benefit from the Boosting algorithm. EP_D is the best choice but its improvement from base classifier is very limited.

It worths to mention that the number of predicted boundaries (noted by ω) from different feature sets varies greatly, for both MAP NB and AdaBoostM1. EP_c has $\omega = 0$ with MAP NB, since the feature set is too simple, and no instance is predicted with a posterior probability higher than 50% for positive class. In AdaBoostM1 training iterations, the 25 false negative classified instances have higher weight out of the 535 population. As a result, the number of positive predictions increase with $\omega = 0.44$. I am delighted to see this effect of AdaBoostM1, but since P_k and WD are mostly unchanged, I believe this prediction result still contains many false positive cases. Similarly, EP_D and EP_{DR} have $\omega > 1$ in MAP NB, which means they have more FP than FN. AdaBoostM1 emphasise on FP predictions and it has the effect to reduce FP in validation set.

Generally, AdaBoostM1 with MAP NB base classifier is capable to improve the ω score, but its P_k and WD are mostly determined by MAP NB (Table 7.27). On the other hand, AdaBoostM1 with C4.5 base classifier improves decision tree performance with almost every metric. Boosting with C4.5 is a better choice than MAP NB.

Bagging works through making multiple versions of models and unweighed plurality vote when predicting a class, where the multiple versions of models are based on bootstrap replicates of the training set. So Bagging is designed to avoid potential overfitting errors. Bagging with MAP NB base classifier (Table 7.29) has no positive effect over MAP NB (Table 7.27) except improving ω on EP_D and EP_{DR}.

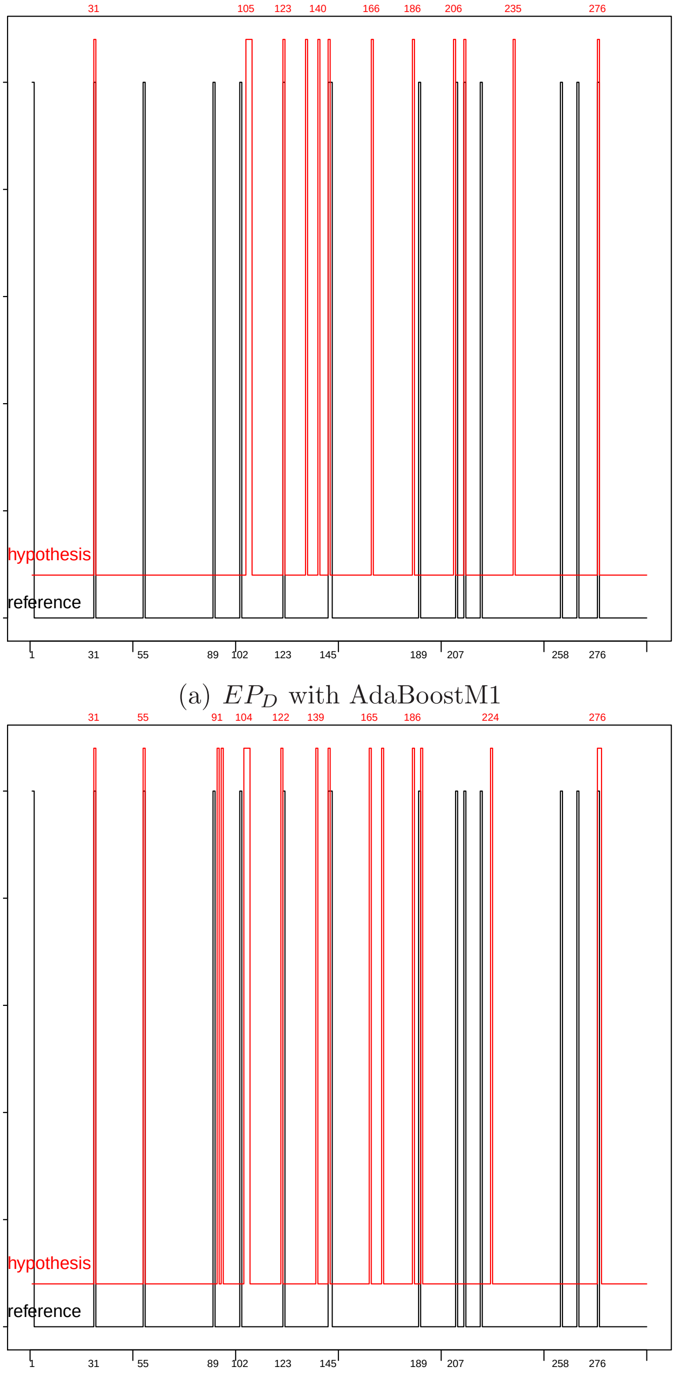
Each bag of instances are sampled with replacement from training set. Since MDTM data set is highly unbalanced with positive instances lower than 1%, the replacement is not likely to highly increase the portion of positive instances. On the other hand, MAP NB is built from maximum likelihood estimation, and it generates robust predictions with inaccurate probability estimation [Domingos and Pazzani, 1996], so the base classifiers of a Bagging model are likely to be similar and stable. I do not expect a big difference appear from MAP NB and a Bagging model. Quite similar P_k and WD scores from Table 7.29 and Table 7.27 validate this assumption and hence naïve Bayes is not a good choice for base classifier of Bagging model. On the contrary, Bagging with C4.5 base classifier makes a difference. A decision tree is constructed with information gain, so different data sets modify decision tree structure and results in different predictions. Table 7.25 and Table 7.26 show explicit difference in segmentation accuracy. Unfortunately, such difference has negative effect for Bagging in MDTM corpus. The decision tree alone has the best segmentation accuracy with EP_D.

The observations from ensembles suggest that Bagging has very limited effect over both simple and horizon based feature sets. Nevertheless, Boosting actively modifies base classifier performance. Boosting with C4.5 base classifier works well with EP_c simple features, and Boosting with MAP NB base classifier works well with EP_D VE horizon features.

7.5 Conclusion

In this chapter, I implemented multiple logistic regression model and several classification models for MDTMs topic segmentation. The regression model is a competitive approach to predict topic boundary but its accuracy depends on an arbitrary threshold for positive class. On the contrary, classification methods avoid arbitrary settings. Thresholded naïve Bayes classifier and Boosting classifiers greatly improve MAP naïve Bayes and C4.5 decision tree separately in PCD segmentation.

First of all, classification unit definition and feature generation are conducted in Section 7.3.1. Vocalisation Event(VE) is used as classification unit, and successful features from AMI segmentation (speaker ID, VE duration) are included



(a) EP_D with AdaBoostM1

(b) EP_{DR} with Bagging

Figure 7.10: Plot of predicted boundaries from winning feature sets in Boosting and Bagging (naïve Bayes base classifier)

in MDTMs experiments. Empty pause, filled pause, overlap and some acoustic features are replaceable or trivial in the AMI corpus, so as to be excluded.

Proper format of features may influence segmentation accuracy. For example, speaker ID is included as a key feature. MDTMs have 22 participants at most, and they can be grouped into 11 roles (Table 7.17). Test shows that speaker role is more effective than speaker ID (Table 7.18), so the latter is no longer used in MDTMs experiments. Vocalisation duration, as a continuous variable, has high skewness (Figure 7.2), but it leads to higher segmentation accuracy than Log transformed duration. A probable explanation is that VE duration owns more variability than Log format, and such variability aids a naïve Bayes classifier.

Furthermore, I extend the concept of Vocalisation Horizon (VH) to both VE duration and speaker roles. Equation 7.14 to 7.17 represent four feature sets to be tested with classifiers. Experiments in this chapter prove the importance of Horizon features on MDTMs segmentation. Speaker role horizon exhibits better accuracy (Table 7.22) than in the AMI corpus (Table 6.16), because MDTMs include 22 speakers on 11 roles, much more than 4 participants in AMI. MDTMs experiments refines the classification approach on content-free topic segmentation.

Naïve Bayes classifier selection and evaluation procedure is in Section 7.4.1. The NB classifier (Table 7.27) results in fair segmentation accuracy and validates the utility of vocalisation horizon on both VE duration and speaker roles. Proportional Thresholds (PT NB) (Table 7.22) further improves NB accuracy and significantly enhances the performance of simple feature set. EP_c and EP_{DR} are of best choice for PT NB.

Other than Bayesian classifiers, ensemble classifiers (Section 7.4.2) are also possible solutions for topic segmentation. Boosting with C4.5 base classifier significantly amends ω score and generates the best accuracy with EP_c features.

As a conclusion, multiple logistic regression and classification models successfully predict PCD boundaries in MDTM corpus. The highly unbalanced data set for classification is a great challenge, but proper setting of probability threshold (PT NB) as well as using ensemble classifiers increase segmentation accuracy. I also find that horizon features of vocalisation event duration and speaker role are essential for classifiers. Topic segmentation with content-free features is promising with experiments in this chapter.

Chapter 8

Speaker Role Identification

In this study I analyse to which extent content-free analysis can capture the turn taking structures of roles, which are commonly seen as intimately related to content. Participant role definitions may be not unique for general meetings. However, the AMI corpus offers unique definitions for each meeting participant, which facilitates our research.

Here I discuss our approaches to classify the vocalization events according to speaker role using content-free measures for analysis. Acoustic features and conversational features are involved, and meeting transcription is not utilised.

8.1 Background

The importance of different types of features, content-free and content-dependent, has been investigated for some meeting analysis tasks [Hsueh and Moore, 2007a]. For instance, they examined the effectiveness of multimodal features on the task of dialogue segmentation within a Maximum Entropy framework. They found that lexical features (i.e. cue words) are the most essential feature class to be combined into the segmentation model. However, lexical features must be combined with other features, in particular conversational features (i.e. lexical cohesion, overlap, pause, speaker change), to train well-performing models. On the other hand, many non-lexical feature classes (e.g. prosody), are not beneficial for recognising dialogue boundaries when used in isolation. Esposito et al. [2007] indicated that pauses are not only generated by psychological motivations but they are also used as a linguistic means for discourse segmentation. Pauses are used by children and adults to mark the clause and paragraph boundaries. Following the research above, I investigate if conversational features (such as pauses and overlaps) are useful for detecting a specific discourse structure and for speaker role.

Banerjee and Rudnicky [2004] conducted fundamental research on meeting state detection and meeting participants role detection, with only speech-based features. They trained a decision tree classifier that learns to detect these states and roles from simple speech-based features that are easy to compute automatically. This classifier detects meeting states 18% more accurately than a random classifier, and detects participant roles 10% more accurately than a majority classifier. The results imply that simple and easy to compute features, e.g. the frequency of speaker change, can be used for this purpose.

In Banerjee and Rudnicky's work, the classifier's input is from a sliding window along a time sequence. They designed a taxonomy of *meeting states*[1]. For the task of meeting states detection, four features are extracted from audio material in the window: (i) the frequency of speaker change within the window of meeting time, (ii) the number of participants who have spoken within the window[2], (iii) the number of overlaps in speech, (iv) the average length of the overlaps. For participant's role detection, more features are included, i.e. the total amount of speech from one speaker in the window, and the amount of overlap speech. To obtain the features above in an automatic way, speaker segmentation and clustering techniques can be used [Reynolds and Rose, 1995]. Thus Banerjee's method is a convenient approach for automatic meeting structure detection, compared with meeting transcription or manual annotation of features. Theoretically, it may be troublesome to detect *overlap* in continuous speech. *Voiced unclassified speech* can be used instead of *overlap*, and then these Voiced unclassified speech events can be labelled with corresponding speakers. The corresponding speakers are the identified speakers of the previous and following vocalizations.

Laskowski et al. [2008] applied new features to the task of role classification (by

[1] Two major meeting states are defined in this paper: discussion and information flow. Moreover, information flow includes two sub-states: presentation and briefing

[2] The length of a window varies from 1 second to 60 seconds. The author used different window length to test the detection accuracy of meeting states and participant roles.

content-free methods): (i) talkspurt initiation, (ii) talkspurt continuation, for (iii) single-participant vocalization, or (iv) vocalization overlap. Talkspurt initiation for vocalization overlap leads to 53% single-feature-type 4-way classification rate on the AMI corpus. I am using similar features as those used by Banerjee and Rudnicky, but in different format and to a more general meeting corpus, the AMI corpus.

8.2 Experiments

8.2.1 Effect of Different Feature Sets

I use 8 meetings[3] from the AMI corpus. In each meeting there are four participants. The project manager (PM), the industrial designer (ID), the user interface designer (UI), and the marketing expert (ME) are labeled as speaker A, B, C and D respectively. For each meeting the feature set holds 990 instances of vocalizations on average. The number of vocalizations in each AMI meeting ranges from 172 to 2014. For each meeting the classification outcome is based on 10-fold cross-validation. Since the target variable Speaker Role is categorical, C4.5, Naïve Bayes and Bayesian Network algorithms are chosen as classifiers. For feature sets, I choose three ways to combine the features, these are VOC_Horizon, GAP_Horizon, and SUM_Horizon, as shown in Table 8.1. VOC_Horizon contains only features related to the present vocalization event and adjacent conversation events. GAP_Horizon contains only features related to the present vocalization event and adjacent pauses and overlaps. SUM_Horizon contains all of the features in these two sets (VOC_Horizon and GAP_Horizon).

Table 8.1 shows three feature sets from one meeting. By comparing the performance of classifiers operating on each of these sets I can distinguish which feature sets are more powerful for speaker role classification. Table 8.2 shows that the Bayes Network classifier performs best for any of the feature sets. The

[3] I may use more meetings in any experiment. In the AMI corpus, 4 meetings are one set, including project introduction, functional design, conceptual design and detailed design (Section 3.1). Moreover, the participants in one set of meetings are fixed. I select 2 sets of meetings arbitrarily when I design tests.

Table 8.1: Three feature set combinations

Feature Set	Features
VOC_Horizon	voc_start, voc_dur, z1_dur, z2_dur, z3_dur, y1_dur, y2_dur, y3_dur
GAP_Horizon	voc_start, voc_dur, PO_z1_dur, PO_z2_dur, PO_z3_dur, PO_y1_dur, PO_y2_dur, PO_y3_dur, PO
SUM_Horizon	voc_start, voc_dur, z1_dur, z2_dur, z3_dur, y1_dur, y2_dur, y3_dur, PO_z1_dur, PO_z2_dur, PO_z3_dur, PO_y1_dur, PO_y2_dur, PO_y3_dur, PO

classification accuracy of 39.54% is approximately 15% higher than the baseline[4]. This result is also higher than the 10% gain reported in Banerjee and Rudnicky [2004]. It should be noted, however, that the data sets used in Banerjee and Rudnicky [2004] are different from ours, as are the definitions of 'roles'. In Banerjee and Rudnicky [2004], there are only 3 roles: presenter, information provider and information consumer, and the speakers for each role vary as the meeting progresses. In our data set, the AMI corpus, there are 4 roles, and each participant's role is fixed throughout the meeting.

Performance details of each classifier are shown in Figure 8.1 and Figure 8.2. The three feature sets perform differently with each classifier, as shown in Table 8.2. C4.5 classifier generates the highest accuracy with VOC_Horizon features, and two Bayesian classifiers generate the highest accuracy with GAP_Horizon features. There is little advantage in using SUM_Horizon even the best performance (39.43%) is less than that achieved by Bayesian Network classifier with GAP_Horizon. I see that, when the classifier is chosen, the performance of different feature sets does not differ much. The choice of classifier appears to be important and in our experiment the Bayesian Network classifier performed best.

[4] In the AMI corpus, only 4 participants are involved in each meeting. Each participant holds a unique role. There are only 4 classes in our study. If all vocalizations are allocated to a random class, the accuracy value is 25%.

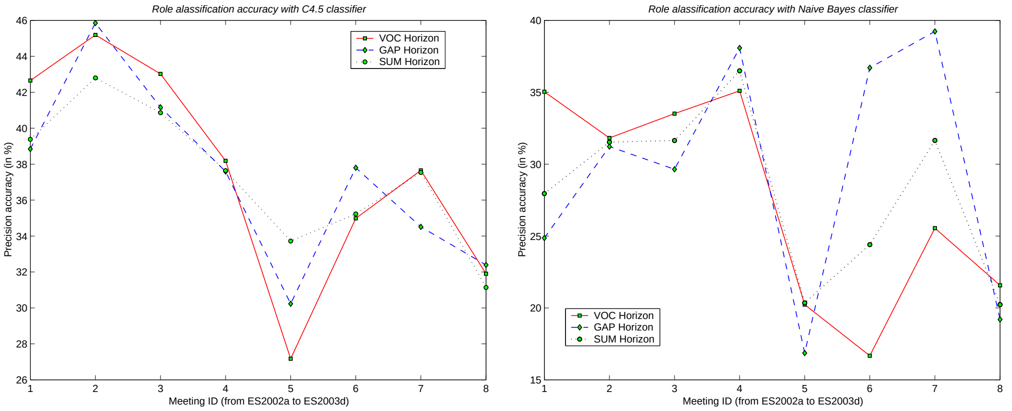


Figure 8.1: C4.5 classifier accuracy on Speaker Role Classification (left) and Naive Bayes classifier accuracy on Speaker Role Classification (right)

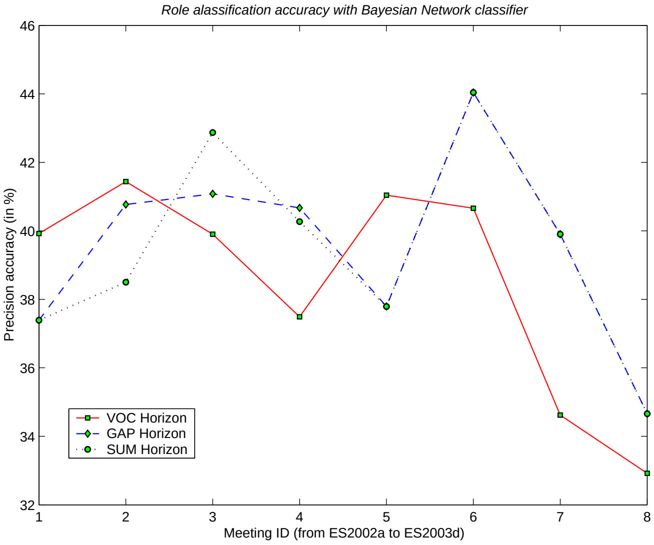
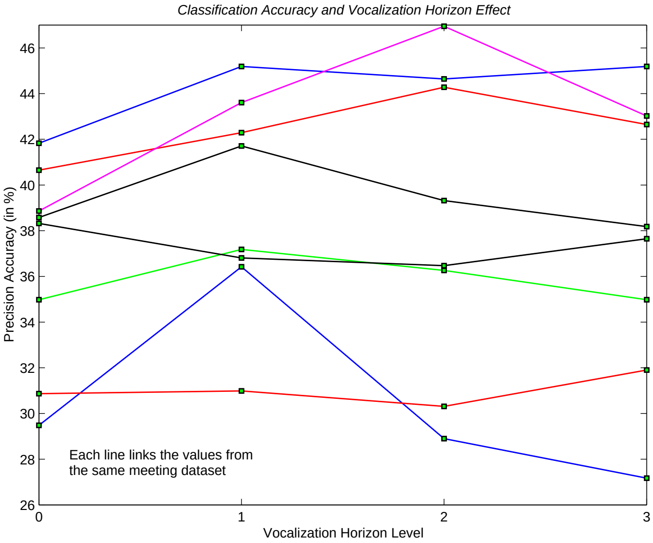


Figure 8.2: Bayesian Network classifier accuracy on Speaker Role Classification

Table 8.2: Classification accuracy from three feature sets and three classifiers

	VOC_Horizon	GAP_Horizon	SUM_Horizon
C4.5	37.59%	37.30%	37.29%
NaiveBayes	27.44%	29.47%	28.03%
BayesNet	38.50%	39.54%	39.43%



Note: each line links the values from same meeting data set. The meetings here are labeled from ES2002a to ES2003d in the AMI corpus.

Figure 8.3: Speaker Role Classification Accuracy and Vocalization Horizon Effect

8.2.2 Effect of Vocalisation Horizon

I follow the hypothesis that consecutive vocalization events influence each other and are not independent. The degree of influence decreases as the distance increases between the current vocalization and the horizon vocalization. In this Section, I describe the experiment to investigate if Vocalization Horizon is useful for role detection in meeting recordings, and to which extent the levels of Vocalization Horizon can influence classification accuracy.

VOC_Horizon feature set is built upon 8 meeting recordings. C4.5 and Bayesian Network classifiers are used for classification. The Bayesian Network classifier obtains the highest accuracy in Table 8.2, so I also test it on Vocalization Horizon features.

The first row of Table 8.3 show that accuracy values of Horizon levels 1, 2, 3 are quite close to each other. The paired t-test shows $p = 0.0805$ between Horizon = 0 and Horizon = 3. For other Horizon levels, the p values are similar. Although statistical significance (at the $p < 0.05$ level) for the effect of VOC_Horizon features on accuracy could not be demonstrated with the Bayesian Network classifier, some influence influence on is still observable. For decision trees (Table 8.3, row 2) I see that the highest mean classification accuracy emerges when Horizon = 1. I find that the accuracy does not improve when I add more vocalization features. The effect of Horizon varies with different data sets. Figure 8.3 shows that in a single meeting the best accuracy may emerge for Horizon = 2. Therefore I do not recommend adopting a fixed optimal Horizon. On the other hand, from Figure 8.3, I see the fact that in most feature sets the classification accuracy is higher for Horizon = 1 than for Horizon = 0. The paired t-test shows $p = 0.023$ between Horizon = 0 and Horizon = 1. The effect of Vocalization Horizon is statistically significant[5].

The results of C4.5 and Bayesian Network classifier show that Vocalization Horizon features improve classification accuracy in general. The improvement from C4.5 is statistically significant, and Hypothesis 2 is supported. I can apply Vocalization Horizon to speaker role detection with confidence.

Table 8.3: Vocalization Horizon effect with VOC_Horizon feature sets on BayesNet and C4.5 classifiers.

	Horizon = 3	Horizon = 2	Horizon = 1	Horizon = 0
BayesNet	38.50%	38.40%	38.48%	37.50%
C4.5	37.59%	38.39%	39.28%	36.70%

8.2.3 Effect of Pauses and Overlap Horizon

I use the Bayesian Network classifier to test the effect of Pause Horizon and Overlap Horizon. As before, the paired t-test is used to evaluate the significance of accuracy difference.

[5] Normality is satisfied for both data sets used in this section.

In Table 8.4, Horizon = 3 gives the highest mean accuracy , but the accuracy difference with Horizon = 0 is not significant. For Horizon = 3, most accuracy values are higher than the ones in Horizon = 0, but in meeting 2002a and 2003a, accuracy decreases. This phenomenon illustrates that classification accuracy does not favour greater Horizon level in all cases, and a fixed Horizon level for all data sets is not recommended. Again in Table 8.4, I find that $p = 0.059$ with Horizon = 1, approaching the 5% significance criterion. I regard this as evidence of the effectiveness of Pause and Overlap Horizon for role classification, though further study is necessary to settle this question conclusively.

Table 8.4: Effect of Vocalization Horizon on a BayesNet with GAP_Horizon feature sets.

MeetingID	Horizon = 3	Horizon = 2	Horizon = 1	Horizon = 0
2002a	37.39%	39.56%	39.38%	39.93%
2002b	40.77%	41.31%	39.51%	38.42%
2002c	41.08%	39.60%	38.34%	37.37%
2002d	40.67%	38.83%	36.69%	34.56%
2003a	37.79%	37.21%	40.70%	41.28%
2003b	44.04%	44.04%	44.77%	42.75%
2003c	39.90%	38.55%	40.57%	34.34%
2003d	34.66%	34.32%	35.80%	33.41%
Mean	39.54%	39.18%	39.47%	37.76%
p-value	0.19	0.185	0.059	

Chapter 9

Evaluation of Content-free Topic Segmentation

In this chapter I review the motivation, methodology, problems, observations and findings in content-free topic segmentation research.

9.1 Migrating from Segmentation to Classification and Regression

There are many segmentation methods implemented to indicate changes in continuous speech, such as speaker segmentation [Chen and Gopalakrishnan, 1998] and speech non-speech segmentation [Jørgensen and Mølgaard, 2006]. These methods evaluate the similarity of continuous speech via selected acoustic properties, and designate the segment boundary where a low similarity score occurs. Topic segmentation, as a text-based approach, has been well studied and has successfully employed lexical cohesion methods [Hearst, 1997]. However, for text independent topic segmentation, the similarity based approach is inadequate. Although I can define a speaker's voice with Gaussian mixture models and calculate coherence by Bayesian Information Criterion (BIC) [Schwarz, 1978], it is difficult to define the acoustic property for a $Topic$. In other words, a person holds stable acoustic characters in his/her voice, and such characters are replicable in speech. However, there is no evidence that a certain topic holds stable acoustic characters in different practices and environments.

I therefore pursued a text independent approach, and transformed topic segmentation into a topic boundary detection task. In this way, the difficulty of defining the properties of an integral topic is eliminated. On the contrary, it is more practical to define two classes: topic boundary VEs and non-boundary VEs. I assume that topic boundary instances have different acoustic and vocalisational properties than in-topic instances. In order to validate this assumption, I presented the detailed design of this binary classification approach in Chapter 4, and following that, the topic boundary detection experiments in the AMI corpus and the MDTM corpus are presented separately in Chapter 6 and 7. Logistic regression is tested with MDTM corpus (Section 7.2).

9.2 The advantages of vocalisational events

Most classification schemes are based on random and independent samples, which are pieces of speech in topic segmentation study. Banerjee and Rudnicky [2004] used a moving window to sample speech features during a meeting, and built a C4.5 decision tree to predict one out of four meeting states. These features include the times of speaker change in window, the frequency of overlaps and the duration of overlaps. Banerjee found that the highest accuracy of detecting meeting states (51.1%) emerges at the window size of 20 seconds.

This moving window design has the advantage of expressing the frequency of speaker change and overlaps, but it cannot extract the properties and patterns from each speaker's voice which contains essential acoustic and vocalisational features for topic boundary classification. Luz and Su [2010] proposed Vocalisation Event(VE) as a classification unit (Section 3.1.2) instead of a fixed length moving window. VE is extracted with speaker segmentation techniques and only contains voice from one speaker. So it is convenient to extract various properties from VE, such as start time, speaker ID, role, duration, pauses, overlaps.

A novel concept Vocalisation Horizon (VH) is employed to express the sequential features from adjacent VEs (Section 3.1.2.4). Simple VE feature set EP (Equation 3.5) achieves topic segmentation accuracy of $P_k = 0.378$ with $\omega = 0.962$ (Table 6.3), which is about 13% better than proportional uniform baseline. Furthermore, horizon based feature set EP_{GAP} (Equation 3.8) achieves the highest segmentation accuracy $P_k = 0.326$ with $\omega = 1.144$ (Table 6.3) in the AMI corpus.

After a comprehensive study of various vocalisation features, I find that empty pauses are more influential than filled pauses (Section 6.3.1). The VOC horizon (VE duration) and GAP horizon (empty pauses and overlaps) have advantage with Bayesian classifiers and ensemble classifiers respectively. So vocalisation event and its features are good data representations for text independent topic segmentation.

9.3 Effectiveness of segmentation metrics

If one classifier can achieve 100% accuracy, the position of topic boundary is perfectly predicted. However, when a False Positive prediction is near to the true boundary, its influence on the task of a user browsing the contents of a meeting is different from a FP prediction far from a real boundary. However, such a FP case has the same influence on classification accuracy. WD and P_k are designed to check the closeness of predicted instances to the reference, but they are not reliable enough to distinguish the effects of inadequate and reduandant topic boundaries predicted[1]. A balance factor ω is introduced (Section 5.3) in order to assess the weakness of under or over prediction (ω is analogous to recall).

Given this evaluation setup, I find that while CRF is always better than PT NB on WD, its ω is too low for it to be useful in AMI topic segmentation (Section 6.2.5.2). The definition of ω further shows that accuracy as well as precision and recall are not appropriate metrics for segmentation. A good segmentation may coexist with low accuracy. I propose WD, P_k and ω together as a unified metric set for segmentation evaluation.

[1] WD and P_k are analogous to precision, which examines the percentage of hit out of all predicted instances, but lacks measure of unpredicted true instances.

9.4 Measures against skewed and correlated instances

Most classifiers are based on independent samples and favor balanced data sets. Our instances for segmentation, on the other hand, are sequential and highly imbalanced. In order to bridge this gap, I propose to adapt both feature sets and classifiers. Feature sets, such as *Vocalisation Horizon* and *GAP Horizon* are employed to represent the sequential character of instances. The Horizon succeeds, at least in part, in integrating broader correlation among instances. In the EPs and FPs feature sets, FP_{VOC} and EP_{VOC}(Equation 3.2 and 3.6) improve WD and P_k over simple features (Equation 3.1 and 3.5) with 3 types of NB classifiers (Table 6.3). However, EP_{VOC} reduces the ω score with PT NB (Table 6.3). When ω is almost 1, the effort to increase precision may cause a classifier to predict less boundaries.

In future work, I propose to generate a novel feature ("distance measure") reflecting the Euclidean distance of current instance to the next possible positive instance, so that the nearer instance is assigned higher value. The motivation of "distance measure" comes from Vocalisation Horizon, where sequence information of confined neighbour instances is coded as a feature. Since sequence information improves segmentation accuracy in many classification trials, I prefer to explore further in this orientation. A prominent fact is that, if an instance is a true boundary, its closely adjacent neighbours have low probability to be a boundary, but such probability rises at a certain distance (possibly the mean duration of a topic in the format of vocalisation numbers). I plan to coin a new feature to represent the probability variation and compare segmentation accuracy with established methods.

Although features can be adapted to accommodate sequential instances and such adaptation has advantages, data imbalance can be adjusted mostly by classifier design. MAP NB predicts boundary better than C4.5 decision trees (Table 4.1), and can be improved through thresholding (Table 6.3). NB suffers less from data imbalance and data dependency. Zhang [2004] explained the condition of NB's optimality as the case when dependencies among attributes cancel each other out. Furthermore, NB allows us to increase the threshold for positive class, and thus mitigate imbalance. Proportional thresholding generates better accuracy than fixed thresholding. On the other hand, CRF disappointed in terms of its ability to model dependent features. Again, imbalance between positive and negative instances may be the culprit. Further research is needed in this area.

Beyond probability based classifiers, ensemble classifiers are promising alternatives for dealing with data imbalance. Boosting with C4.5 base classifier significantly reduces false positive predictions and presents outstanding segmentation accuracy in the AMI corpus. However, Bagging does not show clear advantage over either of its two base classifiers (C4.5 and naïve Bayes). One explanation for this is that the instances for Bagging are sampled with replacement from the training set. This re-sampling procedure is not likely to change the ratio of samples in two classes, and consequently Bagging is not a powerful tool to mitigate the influence of highly unbalanced data set. On the other hand, Boosting performs weighted learning over the whole training set and actively increases the weight on difficult cases, which helps to balance the data set and learn from the less populated positive instances (topic boundary class). Boosting successfully outputs almost balanced predictions and accuracy comparable to PT NB (Table 6.8). If C4.5 decision tree is used as base classifier, the complexity of the tree influences segmentation accuracy.

9.5 AMI corpus and MDTM corpus

The AMI corpus is a series of simulated meetings with predefined scenario and themes, and its topics are defined and extracted after meetings. MDTMs are real medical team meetings arranged by patient cases, where each PCD is treated as one topic. A simple statistics in Table 9.1 shows that in the AMI corpus VE is relatively short and number of VE per topic is nearly 2 times of that in MDTM. So the class imbalance problem in the AMI corpus is much more serious.

With PT NB classifier, AMI and MDTM have similar segmentation accuracy on most feature sets, except on Role Horizon features, MDTM is better (Table 6.16 and 7.22). With the Boosting classifier, MDTM have better accuracy in most cases (see Table 6.9 and 7.24).

Since PCD has been described as a well structured talk [Kane and Luz, 2006],

Table 9.1: Statistics on AMI and MDTM corpus

	Mean duration of VE (sec)	Std. Dev. of VE duration	Avg No. of VE per Topic	Mean duration per Topic (sec)	Number of roles
AMI	3.98	7.35	34.25	139.6	4
MDTM	10.68	17.72	21.52	244	10

Table 9.2: MAP NB classification results in AMI and MDTM corpus

(a) AMI corpus

	P_k	WD	ω
EP	0.408	0.54	1.78
EP_{ROLE}	0.363	0.471	1.22
EP_{VOC}	0.365	0.496	1.5

(b) MDTM corpus

	P_k	WD	ω
EP_c	0.417	0.455	0
EP_R	0.398	0.465	0.96
EP_D	0.272	0.37	1.6

I am interested in how much this structure influences segmentation accuracy. The most fundamental structure is PCD, and the turns of speaker roles are relatively stable within a PCD, so there is assumption that role turns indicate PCD boundaries. Among all feature sets in this study, Role Horizon best represents role turn. I revisit AMI and MDTM experiments, and list segmentation accuracy of MAP NB classifier in Table 9.2. The reason to use MAP NB instead of PT NB is that MAP NB defines a 50% probability threshold for positive class. Different feature sets are comparable under this fixed threshold instead of proportional thresholds. EP and EP_c refer to the same features in the AMI corpus and the MDTM corpus separately, so does EP_{ROLE} and EP_R, EP_{VOC} and EP_D. In Table 9.2(b) I see that with EP_c MAP NB predicts that no instance has probability higher than 50% in topic boundary class, but EP_R significantly improves it and generates $\omega \approx 1$, which is better than EP_D. This observation suggests the power of internal meeting structure on boundary prediction. On the other hand, Role Horizon effect of the AMI corpus is less significant (Table 9.2(a)). When I separate the AMI data set into four groups and each group contains the meetings from the same project phase, the segmentation accuracy increases (Section 6.4.1). So the homogeneity of meeting content also influences segmentation result. In future work, I would like to probe more meeting structure related features to indicate topic boundary.

9.6 Content-free segmentation and text dependent approaches

In this section, I compare our segmentation results with other well known methods. These methods include the lexical cohesion based LCSeg model [Galley et al., 2003] and Maximum Entropy model with comprehensive features [Hsueh and Moore, 2007a]. Table 9.3 lists the best results from each model and the ω score of MAXENT is transformed from the reported number of hypothesis boundaries.

Table 9.3: Comparison with other meeting segmentation methods. Only the best reported results are presented.

Method	Corpus	Segm. level	P_k	WD	ω
LCSeg [1]	ICSI	top-level	31.91%	35.88%	-
LCSeg [2]	AMI	top-level	40.00%	49.00%	-
MAXENT [2]	AMI	top-level	30.00%	33.00%	0.89
LCSeg [2]	AMI	sub-topic	40.00%	47.00%	-
MAXENT [2]	AMI	sub-topic	34.00%	36.00%	0.52
PTNB+EP_{GAP}	AMI	sub-topic	34.4%	42.9%	0.83
AdaBoostM1+EP_c[3]	MDTM	top-level	25.6%	31.4%	0.92

[1] Galley et al. [2003], [2] Hsueh and Moore [2007a], [3] with C4.5 base classifier

LCSeg is the most popular topic segmentation model for text, and it presents leading P_k and WD score in ICSI corpus. But LCSeg performs much worse in the AMI corpus. A number of factors may account for this. On average, ICSI meetings last one hour, with 7 top-level topics and 10 more sub-topics. AMI meetings are about half hour long with 8 top-level topics and 3 more sub-topics. The semantic structure of the meetings is also important. AMI meetings are composed of "agenda-based conversation segments" (e.g., presentation, group discussion) that are typically signalled by differences in group activity [Hsueh and Moore, 2006]. This comparison indicates the weakness of LCSeg and the need of meta features (such as speaker role).

MAXENT shows the best segmentation accuracy in the AMI corpus, which indicates the success of lexical features combined with conversational features, prosody features, etc. In the AMI corpus, the accuracy of MAXENT is much better than LCSeg on either top-level topics or sub-topics. This observation verifies the practicality of using conversational features to predict topic boundary, although keywords and lexical cohesion scores are also important features.

In order to process content sensitive meetings and avoid ASR errors, I investigated content-free methods and prefered to use only text independent features to detect topic boundaries in meetings. CONV features were tested to be less competitive than ALL features in the AMI corpus [Hsueh and Moore, 2007a], but in case of P_k and WD, the segmentation accuracy of CONV and ALL have only minor difference. I managed to improve CONV performance in three ways. First, to include more vocalisation features and modify them into more suitable formats. As introduced in Section 3.1.2.4, $VocalisationHorizon$ is used to incorporate time series information in classification. I explored VE duration, empty pause, overlap and speaker role in horizon format, and found that they are competitive with different classifiers. Horizon remarkably improves topic segmentation accuracy. Second, various classification schemes were studied in detail to minimise the influence of imbalanced data set in two classes and to accommodate the dependency of instances. Third, I evaluated segmentation accuracy with a set of robust metrics. I introduced a new metric ω, which penalises the classifiers predicting very few boundaries and hence guide the study toward high fitness of segmentation.

With these efforts, PTNB+EP_{GAP} model has very similar P_k score as MAXENT in sub-topic segmentation. Although the WD of PTNB is about 7% behind MAXENT, PTNB is still a competitive alternative of MAXENT with ALL features. The difference in VE sampling and *spurts* in Hsueh's work may partially explain the WD scores. More importantly, PTNB has a much better ω score than MAXENT, which may better satisfy meeting browser's needs.

In MDTM, AdaBoostM1 with C4.5 base classifier generates the best P_k, WD and ω in scope. Although direct comparison is not recommended due to different algorithm settings, the outstanding segmentation accuracy of Boosting is encouraging. The locations of prediction boundaries match the reference boundaries well (Figure 7.9(a)). Boosting performs weighted learning over the whole training set, and increases the weight on difficult cases. Highly unbalanced data sets benefit from Boosting because easy cases (mostly in negative class) have low weight.

The successful segmentation practice in MDTM validates our composite design on both feature set modification and classifier selection. Moreover, the relatively stable structures of MDTM and PCD may also account for Boosting's performance. Content-free topic segmentation models are promising retrieval methods for real meetings.

Chapter 10

Conclusion

In this research, I successfully conduct content-free methods for meeting topic segmentation. This is the first time that completely text independent methods achieve outstanding segmentation accuracy on both simulated experimental corpus and real meeting recordings. This research bridges the gap between vocalisation patterns and meeting contents, and consequently I can analyse pragmatic structure of meetings without acknowledgement of contents. For confidential meetings, their content is secured with content-free settings. Finally I deliver a set of robust models to predict topic references for meeting recordings. Meeting audience then has non-linear access to linear recordings.

10.1 Achievements and innovations

In this section, I present the most important observations of our research, and discuss them with respect to the "state of the art" in meeting segmentation.

1. Topic boundary detection is achieved through robust classification schemes.

   In recent topic segmentation research, classification models are widely used to distinguish topic boundary instances from other instances, but people mostly emphasise features and overlook classifier selection. C4.5 decision tree [Banerjee and Rudnicky, 2004] and Maximum Entropy classifier [Hsueh and Moore, 2007a] are employed as the only choices in their work respectively. I investigate classifier selection in both the AMI corpus and the

MDTM corpus and find that the unbalanced data set is the main challenge for commonly used classifiers. The adapted proportional threshold naïve Bayes classifier and Boosting classifier are robust with proper combination of vocalisational features. Based on successful classification schemes, content-free topic segmentation is competitive against text dependent approaches[1].

2. Vocalisation Event (VE) instead of fixed length analysis window is used as classification unit (instance).

   The former has two advantages with respect to the latter one in meeting recording segmentation. First, VE naturally accommodates vocalisational features, such as speaker change, preceding and subsequent pauses, speaker role. Second, VE can be located from audio recordings with speaker segmentation techniques (Chen and Gopalakrishnan [1998]). So VE extraction is independent of transcription or speech recognition. This setting satisfies content-free segmentation requirements. Transcription based topic segmentation methods mostly employ *word* as a basic unit for algorithms (i.e., Galley et al. [2003]). Hsueh and Moore [2007a] used a similar unit *spurt* in Maximum Entropy classifier and achieved good segmentation accuracy. I expect that the use of VE will become more widespread with the advances of meeting retrieval.

3. Features selection is essential for content-free topic segmentation.

   Vocalisation Horizon (VH), as a novel feature concept to indicate temporal/ sequence information in classification instances, was found to increase segmentation accuracy. VE duration, empty pause, overlap and speaker role are features that can be readily used with a VH representation. The performance of VH features is dependent on classifier selection. In many cases, VH related feature sets outperform simple vocalisation features.

4. A set of metrics was proposed to evaluate segmentation fitness.

   Although many topic segmentation approaches are based on classification,

[1] based on the comparisons in Table 9.3

precision and recall are suitable metrics for segmentation evaluation. P_k and WD instead are commonly used [Beeferman et al., 1999], [Pevzner and Hearst, 2002]. However, I found that P_k and WD do not suffice to indicate the quantity of predicted boundaries. Neglecting this quantity seriously biases evaluation decisions. I proposed a supplemental metric ω to gauge the ratio of predicted and reference boundaries. $\omega = 1$ means the quantity is the same. $\omega \to \infty$ and $\omega \to 0$ emerge when predictions are too redundant or too few. ω together with P_k and WD support more reliable judgements of segmentation goodness.

5. Topic boundaries from a better structured meeting are easier to be retrieved.

   As regards the relation between meeting content and topic boundary detection, I find that topic segmentation is more accurate in data sets with relatively homogenous content than on mixed data sets (Section 6.4.1). This observation also suggests a relation between meeting content and conversational features with respect to boundary detection.

10.2 Limitations and future work

While successful content-free topic segmentation strategies have been identified in this research, I also find inefficient aspects and limitations of the current setting, which I plan to address in future research. These issues are summarised below.

1. CRF performs poorly in the AMI corpus.

   In the AMI corpus instances are sampled sequentially and are unbalanced in topic boundary class and non-boundary class. Since CRF accounts for dependency among adjacent observations (class label) where the posterior probability of an observation is only determined by its previous observation and the whole feature sequence through a feature function, I assumed that CRF could perform better on topic boundary detection than other classifiers which assume independent samples. However, CRF predicts very low ω (Section 6.2.5.2). The reason is that its feature function (Equation 4.11) has low generality on positive instances, since very few instances are positive in training set. As a conclusion, CRF is not appropriate for highly unbalanced data, and alternative methods are required to present sequence information among instances.

2. Prosodic features are not fully utilised.

   Speech rate is an essential prosodic feature. Hsueh and Moore [2007a] used the number of words and the number of syllables per second to represent speech rate. Since ASR is not employed in our method, the start and end time of a word is not recognised. It is difficult to extract speech rate from a vocalisation event. I may extract syllables or use energy variation to estimate speech rate in future research. Moreover, speaker based prosodic analysis may be a more appropriate approach, because inter-personal variation is excluded and the patterns of intra-personal prosody variation is more prominent.

3. Vocalisation horizon is the only successful practice to integrate sequence information of instances in classification.

   Such sequence information implies the Euclidean distance of one instance to the nearest boundary. I am interested in exploring methods to explicitly express the distance to topic boundary (note that a simple feature of the distance to boundary is not available for test set). Since classifiers are trained with likelihood or information gain of features, and the likelihood is coded as the frequency of instances with respect to classes, I suggest a likelihood representation based on the distance of instances to the nearest reference boundary.

4. An extrinsic measure is needed to evaluate segmentation.

   P_k, WD and ω are intrinsic evaluation metrics. They tell the goodness of match between predictions and references, but I do not know how well these metrics to extrinsic measures such as task completion and user satisfaction in real meeting browsing tasks. I will survey user feedback on automatically segmented recordings and exploit the factors that affect user experience. Consequently, such factors will be used to adjust classification models.

Bibliography

Alan Agresti. *Categorical Data Analysis.* John Wiley, New York, 2nd edition, 2002.

Satanjeev Banerjee and Alexander Rudnicky. Using simple speech-based features to detect the state of a meeting and the roles of the meeting participants. In *INTERSPEECH-2004*, pages 2189–2192, 2004.

Satanjeev Banerjee, Carolyn Rose, and Alexander I. Rudnicky. 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)*, pages 643–656, 2005. ISBN 978-3-540-28943-2.

Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999. ISSN 0885-6125.

Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, USA, 1 edition, 1996.

Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer (version 5.1.05). 2009. `http://www.praat.org/`, accessed April 2011.

Matt-M. Bouamrane and Saturnino Luz. Meeting browsing. *Multimedia Systems*, 12(4–5):439–457, 2007.

Leo Breiman. Bagging predictors. In *Machine Learning*, pages 123–140, 1996.

G. Brown and G. Yule. *Discourse Analysis.* Cambridge University Press, Cambridge, MA, 1983.

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

Scott S. Chen and P. S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

Heidi Christensen, Balakrishna Kolluru, Yoshihiko Gotoh, and Steve Renals. Maximum entropy segmentation of broadcast news. In *in Proceedings of ICASSP 2005*, 2005.

Herbert H. Clark. Managing problems in speaking. *Speech Commun.*, 15(3-4): 243–250, 1994. ISSN 0167-6393.

James M. Jr. Dabbs and Barry Ruback. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Social Psychology*, 20(123–169), 1987.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

Pedro Domingos and Michael Pazzani. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *Machine Learning*, pages 105–112. Morgan Kaufmann, 1996.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pages 973–978, 2001.

Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics Speech and Signal Processing*, 32(6):1109–1121, 1984.

Anna Esposito, Vojtech Stejskal, Zdenek Smekal, and Nikolaos Bourbakis. The significance of empty speech pauses: Cognitive and algorithmic issues. In *Ad-*

vances in Brain, Vision, and Artificial Intelligence, volume 4729/2007, pages 542 – 554. Springer Berlin / Heidelberg, 2007.

Anna Esposito, Vojtech Stejskal, and Zdenek Smékal. Cognitive role of speech pauses and algorithmic considerations for their processing. *IJPRAI*, 22(5): 1073–1088, 2008.

Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27:861–874, June 2006.

Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. for Artif. Intel.*, 14(5):771–780, 1999.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of online learning and an application to boosting,. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 562–569, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Alexander Gruenstein, John Niekrasz, and Matthew Purver. Meeting structure annotation: Data and tools. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 117–127, 2005.

M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.

Gillian Hardstone, Mark Hartswood, Rob Proctor, Roger Slack, and Alex Voss. Supporting informality: Team working and integrated care records. In *CSCW Chicago 2004*. ACM, ACM Press, 2004.

Mark Hartswood, Rob Proctor, Mark Rouncefield, and Roger Slack. Making a case in medical work: Implications for the electronic patient record. *Computer Supported Co-operative Work (CSCW)*, 12:241 – 266, 2003.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.

Michael Jerome Hawley. *Structure out of sound*. PhD thesis, Cambridge, MA, USA, 1993. UMI Order No. not available.

Marti A. Hearst. Multi-paragraph segmentation of expository text. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 9–16, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

Marti A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, 1997. ISSN 0891-2017.

Julia Hirschberg and Christine H. Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of the 5th International Conference on Spoken Language Processing*, page paper0976, 1998.

David W. Hosmer and Stanley Lemeshow. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, 9(10):1043–1069, 1980.

David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. John Wiley and Sons, New York, 2 edition, 2000.

Pei-Yun Hsueh and Johanna D. Moore. Automatic topic segmentation and labeling in multiparty dialogue. In *Proceedings of the first IEEE/ACM workshop on Spoken Language Technology (SLT)*, 2006. AMI-203.

Pei-Yun Hsueh and Johanna D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the 45th Annual Meeting of the ACL*. Association for Computational Linguistics, 2007a.

Pei-Yun Hsueh, Johanna Moore, and Steve Renals. Automatic segmentation of multiparty dialogue. In *Proc. EACL*, 2006.

Peiyun Hsueh and Johanna D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *ACL*. The Association for Computer Linguistics, 2007b.

Rongqing Huang and John H. L. Hansen. High-level feature weighted gmm network for audio stream classification. In *INTERSPEECH-2004*, pages 1061–1064, 2004.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001. ISBN 0130226165. Foreword By-Raj Reddy.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. The icsi meeting corpus. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP '03). 2003*, volume 1, pages 364–367, 2003.

K. W. Jørgensen and L. L. Mølgaard. Tools for automatic audio indexing. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2006. Supervised by Prof. Lars Kai Hansen, IMM.

Jean-Claude Junqua and Jean-Paul Haton. *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. Kluwer Academic Publishers, Norwell, MA, USA, 1995. ISBN 0792396464.

Bridget Kane and Saturnino Luz. Multidisciplinary medical team meetings: An analysis of collaborative working with special attention to timing and teleconferencing. *Computer Supported Co-operative Work (CSCW)*, 15(5-6):501 – 535, December 2006.

Bridget Kane and Saturnino Luz. Achieving diagnosis by consensus. *Computer Supported Cooperative Work (CSCW)*, 18(4):357–391, 2009.

Bridget Kane, Saturnino Luz, Gerard Menezes, and Donal P Hollywood. Enabling change in healthcare structures through teleconferencing. In *Proceedings of* *the 18th IEEE International Symposium on Computer-Based Medical Systems*, pages 76–81, Dublin, Ireland, 2005. IEEE.

Bridget T. Kane. *An analysis of multidisciplinary medical team meeting and the use of communication technology.* PhD thesis, University of Dublin, Trinity College, 2008.

L. I. Kuncheva, M. Skurichina, and R.P.W. Duin. An experimental study on diversity for bagging and boosting with linear classifiers. *Information Fusion*, 3:245–258, 2002.

M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models.* McGraw Hill, 5th edition, 2005.

Peter Ladefoged. *A course in phonetics.* Harcourt Brace College, Fort Worth, third edition edition, 1993.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, 2001.

Kornel Laskowski, Mari Ostendorf, and Tanja Schultz. Modeling vocal interaction for text independent participant characterization in multi-party conversation. In *SIGDIAL2008*, 2008.

Gina-Anne Levow. Prosodic cues to discourse segment boundaries in human-computer dialogue. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, page 93–96, Cambridge, Massachusetts, USA, April 30 - May 1 2004. Association for Computational Linguistics, Association for Computational Linguistics.

Lie Lu and Hong-Jiang Zhang. Unsupervised speaker segmentation and tracking in real-time audio content analysis. *Multimedia Systems*, Volume 10:332–343, 2005.

Saturnino Luz. Locating case discussion segments in recorded medical team meetings. In *SSCS '09: Proceedings of the ACM Multimedia Workshop on Search-*

ing Spontaneous Conversational Speech, pages 21–30, Beijing, China, October 2009. ACM Press. See also Luz [2012].

Saturnino Luz. The non-verbal structure of patient case discussions in multidisciplinary medical team meetings. *ACM Transactions on Information Systems*, 30(4):17, 2012. To appear.

Saturnino Luz and Masood Masoodian. A mobile system for non-linear access to time-based data. In *Proceedings of Advanced Visual Interfaces AVI'04*, pages 454–457. ACM Press, 2004. ISBN 1-58113-867-9.

Saturnino Luz and David M. Roy. Meeting browser: A system for visualising and accessing audio in multicast meetings. In *Proceedings of the International Workshop on Multimedia Signal Processing*, pages 489–494. IEEE Signal Processing Society, September 1999.

Saturnino Luz and Jing Su. The relevance of timing, pauses and overlaps in dialogues: Detecting topic changes in scenario based meetings. In *Proceedings of Interspeech 2010*, 2010.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48, 1991. ISSN 0891-2017.

John Neter, Michael H. Kutner, Christopher J. Nachtsheim, and William Wasserman. *Applied Linear Statistical Models*. McGraw Hill, 4th edition, 1996.

Alan V. Oppenheim. Speech spectrograms using the fast fourier transform. *Spectrum, IEEE*, 7(8):57 –62, August 1970. ISSN 0018-9235.

John Øvretveit. System negligence is at the root of medical error. *International Journal of Health Care Quality Assurance*, 13(3):103–105, March 2000.

L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

Foster Provost and Tom Fawcett. Robust classification systems for imprecise environments. In *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 706–713. AAAI Press, 1998.

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993. ISBN 0-13-015157-2.

Javier Ramirez, Jose C Segura, Carmen Benitez, Angel de la Torre, and Antonio Rubio. Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, 42(3-4):271–287, 2004.

Steve Renals, Thomas Hain, and Herve Bourlard. Recognition and interpretation of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '07)*, 2007.

D.A. Reynolds and R.C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1):72–83, 1995. ISSN 1063-6676.

Klaus Ries. Segmenting conversations by topic, initiative and style. In *Proceedings of ACM SIGIR'01 Workshop on Information Retrieval Techniques for Speech Applications*, 2001.

David M. Roy and Saturnino Luz. Audio meeting history tool: Interactive graphical user-support for virtual audio meetings. In *Proceedings of the ESCA workshop: Accessing information in spoken audio*, pages 107–110. Cambridge University, April 1999.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523, 1988.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Gükhan Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Commun.*, 32(1-2):127–154, 2000. ISSN 0167-6393.

V.L. Smith and H.H. Clark. On the course of answering questions. *Journal of Memory and Language*, 32:25–38, 1993.

Jongseo Sohn, Student Member, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Process. Lett*, 6:1–3, 1999.

A.-B. Stenstrom. *Pauses in monologue and dialogue.* In The London-Lund Corpus of Spoken English: Description and Research. Lund:Lund University Press, 1990.

Marc Swerts, Anne Wichmann, and Robbert-Jan Beun. Filled pauses as markers of discourse structure. In *Proceedings of the 4th International Conference on Spoken Language Processing*, pages 1033–1036, 1996.

B. Widrow, Jr. J.R. Glover, J.M. McCool, J. Kaunitz, C.S. Williams, R.H. Hearn, J.R. Zeidler, Jr. E. Dong, and R.C. Goodlin. Adaptive noise cancelling: Principles and applications. In *Proceedings of the IEEE*, volume 63(12), pages 1692–1716, December 1975.

S. Young. *The HTK Book Version 3.4.* Cambridge University Engineering Department, 2007. URL `http://htk.eng.cam.ac.uk/docs/docs.shtml`. accessed April 2011.

Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference.* AAAI Press, 2004.

Xiaojin Zhu. Conditional random fields. Website, 2010. `http://pages.cs.wisc.edu/~jerryzhu/cs769/CRF.pdf`, accessed April 2011.